



APLICAÇÃO DE TÉCNICAS DE
REDUÇÃO DE DIMENSIONALIDADE
PARA DETERMINAÇÃO DE
CONDIÇÕES OPERACIONAIS ÓTIMAS
EM UM PROCESSO INDUSTRIAL

Paula Araujo Gomes

Projeto de Final de Curso

Orientador:

Prof. Bruno Didier Olivier Capron, D.Sc.

Agosto de 2021

APLICAÇÃO DE TÉCNICAS DE REDUÇÃO DE DIMENSIONALIDADE PARA DETERMINAÇÃO DE CONDIÇÕES OPERACIONAIS ÓTIMAS EM UM PROCESSO INDUSTRIAL

Paula Araujo Gomes

Projeto de Final de Curso apresentado à Escola de Química, como parte dos requisitos necessários à obtenção do grau de Engenheira Química.

Aprovado por:

Maurício Bezerra de Souza Junior, D.Sc.

Aline Barbosa Alves, B.Sc.

Orientado por:

Bruno Didier Olivier Capron, D.Sc.

Rio de Janeiro, RJ – Brasil

Agosto de 2021

Ficha Catalográfica

Gomes, Paula Araujo.

Aplicação de técnicas de redução de dimensionalidade para determinação de condições operacionais ótimas em um processo industrial / Paula Araujo Gomes. Rio de Janeiro: UFRJ/EQ 2021.

x, 69 p.; il.

(Monografia) – Universidade Federal do Rio de Janeiro, Escola de Química, 2021.

Orientador: Bruno Didier Olivier Capron.

1. Machine Learning. 2. Controle de Processos. 3. Redução de dimensionalidade 4. PCA 5. FDA 6. Monografia (Graduação – UFRJ/EQ). 7. Bruno Capron. I. Aplicação de técnicas de redução de dimensionalidade para determinação de condições operacionais ótimas em um processo industrial.

A todas as famílias das vítimas do Coronavírus e aos
profissionais de saúde que desempenham papel de
heróis nesse momento de crise mundial.

AGRADECIMENTOS

À minha mãe que sempre incentivou o estudo na minha vida e em meio a todas as dificuldades existentes por ser mãe solteira de quatro filhos conseguiu garantir um bom futuro para todos.

A todos os professores que tive a honra de passar, não só durante a graduação, mas durante toda minha formação acadêmica que com seus conhecimentos permitiram eu ser o que sou hoje. Gostaria de fazer um agradecimento especial ao meu orientador Bruno Didier Olivier Capron pelo suporte neste projeto e a toda equipe da Preditiva Digital da Radix por me permitirem usar dados fundamentais para a realização deste projeto. Em particular gostaria de agradecer a Aline Barbosa Alves, Ícaro Augusto Maccari Zelioli, José Guilherme Tavares Monteiro e Neander Alessandro da Silva Paula por todo apoio e incentivo que me deram para encontrar um tema para meu projeto final de curso.

Não posso esquecer de mencionar todos os colegas que fizeram parte do meu percurso na graduação tanto no Brasil quanto na França. Em particular agradeço a Bruno Belucci por estar ao meu lado desde que o conheci. Dos meus amigos que conheci no Brasil agradeço a Aluan Ferreira, Breno Mourão, Bruna Conrado, Flávio Fortes, Gabriel Santos, Larissa Silva, Pedro Henrique Cerqueira e Suzana Nunes. Dos meus amigos que conheci na França agradeço pela parceria em meio ao desconhecido a Amanda Martins, Gabriel Callado, Juan Quintero, Kaleb Roncatti, Marcelo Orlando e Maria Clara Novaes.

Resumo do Projeto de Final de Curso apresentado à Escola de Química como parte dos requisitos necessários para obtenção do grau de Engenheira Química

APLICAÇÃO DE TÉCNICAS DE REDUÇÃO DE DIMENSIONALIDADE PARA DETERMINAÇÃO DE CONDIÇÕES OPERACIONAIS ÓTIMAS EM UM PROCESSO INDUSTRIAL

Paula Araujo Gomes

Agosto, 2021

Orientador: Bruno Didier Olivier Capron, D.Sc.

Com a quarta revolução industrial, a aquisição de informações acerca de um processo é feita de uma forma jamais antes vista, onde são obtidos dados em alto volume, alta velocidade e grande variedade. Isso possibilita a aplicação de análise de dados e *Machine Learning* para se obter uma grande gama de informações úteis para a indústria, que podem permitir o aumento da produtividade, da qualidade dos produtos, da segurança dos processos e do lucro das empresas. Uma aplicação popular de *Machine Learning* no controle de processos é a Detecção e Diagnóstico de Falhas (FDD), que possibilita a manutenção de condições operacionais desejadas e previne riscos no processo. O presente trabalho teve como objetivo estudar um processo de produção de resinas poliméricas em um equipamento industrial, visando à determinação de condições operacionais que minimizem a quantidade de imperfeições presentes no produto final. Grandes quantidades de imperfeições em uma resina polimérica levam à formação do chamado produto *Off-spec*, esse produto, se comercializado, terá preço de venda reduzido por apresentar qualidade inferior. Portanto, é desejável a garantia da produção de polímeros com baixa quantidade de imperfeições. Devido ao grande número de variáveis de processo existentes neste projeto, duas técnicas diferentes de redução de dimensionalidade foram usadas: Análise de Componente Principal (PCA) e Análise de Discriminante de Fisher (FDA). Combinando o método de agrupamento *KMeans* e a formação de uma elipse de confiança, regiões operacionais desejadas foram determinadas. Este trabalho apresenta uma comparação entre dois modelos desenvolvidos, o primeiro via PCA e o segundo via FDA. Os resultados obtidos indicaram que, para o problema em questão, ambos os métodos desenvolvidos apresentam desempenho similar. Com esses modelos, foi possível a determinação de *ranges* das variáveis de processo que possibilitam a minimização da formação de imperfeições e, logo, um aumento da qualidade do produto final e do lucro do processo.

ÍNDICE

1	Introdução	1
1.1	Contexto	1
1.2	Motivação	2
1.3	Objetivos	3
2	Revisão Bibliográfica	4
2.1	<i>Machine Learning</i> na indústria	4
2.2	Detecção e Diagnóstico de Falhas	6
2.2.1	Métodos Baseados em Modelos	7
2.2.1.1	<i>Redundância Analítica</i>	7
2.2.2	Métodos Baseados em Dados	10
2.2.2.1	<i>Estatística T^2 de Hotelling</i>	11
2.2.2.2	<i>Estatística Q</i>	12
2.2.2.3	<i>Análise de Componentes Principais - PCA</i>	13
2.2.2.4	<i>Mínimos Quadrados Parciais - PLS</i>	13
2.2.2.5	<i>Análise de Componentes Independentes - ICA</i>	13
2.2.2.6	<i>Análise de Discriminante de Fisher - FDA</i>	14
2.2.3	Métodos de clusterização	14
2.3	Seleção de Variáveis	15
2.4	Importância das Variáveis	16
3	Metodologia	18
3.1	Sistema	18
3.2	Recursos Utilizados	20
3.3	Processamento dos dados	21
3.3.1	Limpeza dos dados	21
3.3.2	Redimensionamento de variáveis	22
3.3.3	Seleção de Variáveis	22
3.3.4	Separação de Dados em Treino e Teste	24

3.4	Modelos de Identificação de Regiões que Minimizam Imperfeições	24
3.4.1	PCA	26
3.4.1.1	<i>Ajuste do modelo</i>	26
3.4.1.2	<i>Transformação dos dados</i>	27
3.4.2	FDA	27
3.4.2.1	<i>Ajuste do modelo</i>	28
3.4.2.2	<i>Transformação dos dados</i>	29
3.4.3	Região Ótima	29
3.4.3.1	<i>KMeans</i>	30
3.4.3.2	<i>Elipse de Confiança</i>	30
3.5	Importância das Variáveis	31
4	Resultados e Discussões	33
4.1	Seleção de Variáveis	33
4.2	Determinação de regiões ótimas via PCA	36
4.2.1	Conjunto de Treino	36
4.2.2	Conjunto de Teste	40
4.3	Determinação de regiões ótimas via FDA	41
4.3.1	Conjunto de Treino	41
4.3.2	Conjunto de Teste	43
4.4	Comparação dos modelos	45
5	Conclusão	49
	REFERÊNCIAS	51
	ANEXO A	56
	ANEXO B	63

ÍNDICE DE FIGURAS

1.1	Fluxo de dados na Indústria 4.0	1
2.1	As seis etapas de um processo CRISP-DM	5
2.2	Classificação de Métodos de FDD	6
2.3	Esquema da detecção de falhas baseada em modelo	8
2.4	Esquema da detecção de falhas por equações de paridade	9
3.1	Estrutura do método de detecção de regiões operacionais que minimizam a quantidade de imperfeições utilizada nesse projeto	25
3.2	Elipse de confiança	31
4.1	Correlação das variáveis de processos com a variável de quantidade de imper- feições	34
4.2	Mapa de calor da correlação das <i>TAGs</i> pré-selecionadas	35
4.3	PCA - Variância Explicada e Acumulada das Componentes Principais	37
4.4	Importâncias das Variáveis sobre as Componentes da PCA	37
4.5	Gráfico 2D das componentes principais de PCA do conjunto de treino	38
4.6	Gráfico 2D das componentes principais de PCA do conjunto de treino com elipse de confiança	39
4.7	Gráfico 2D das componentes principais de PCA do conjunto de teste com elipse de confiança	40
4.8	Importâncias das Variáveis sobre as Componentes da FDA	41
4.9	Gráfico 2D das componentes principais de FDA do conjunto de treino	42
4.10	Gráfico 2D das componentes principais de FDA do conjunto de treino com elipse de confiança	43
4.11	Gráfico 2D das componentes principais de FDA do conjunto de teste com elipse de confiança	44
4.12	Elenco de informações contidas no <i>boxplot</i>	46
4.13	Contribuição das variáveis nos modelos desenvolvidos	48

ÍNDICE DE TABELAS

3.1	<i>TAGs</i> do processo	18
3.3	<i>TAGs</i> de qualidade	20
3.4	Tabela com as versões de cada uma das ferramentas	21
4.1	<i>TAGs</i> selecionadas pelo método de Correlação	35
4.3	Quantidade média de imperfeições em cada <i>cluster</i> determinado por <i>KMeans</i> a partir do modelo via PCA	39
4.4	Quantidade média de imperfeições em cada <i>cluster</i> determinado por <i>KMeans</i> a partir do modelo via FDA	42

LISTA DE ABREVIACOES

AI	<i>Artificial Intelligence</i> , Inteligencia Artificial
BSS	<i>Blind Source Separation</i> , Separao de Fontes Cegas
CRISP-DM	<i>Cross Industry Standard Process for Data Mining</i> , Processo Padro Inter-Indstrias para Minerao de Dados
FDA	<i>Fisher Discriminant Analysis</i> , Anlise de Discriminante de Fisher
FDD	<i>Fault Detection and Diagnosis</i> , Deteco e Diagnstico de Falhas
ICA	<i>Independent Component Analysis</i> , Anlise de Componentes Independentes
IQR	Intervalo Interquartlico
PCA	<i>Principal Component Analysis</i> , Anlise de Componente Principal
PLS	<i>Partial Least Squares</i> , Mnimos Quadrados Parciais
SPE	<i>Squared Prediction Error</i> , Erro de Previso ao Quadrado
TAG	Etiqueta de Identificao

1 Introdução

Este capítulo começa apresentando os aspectos gerais da utilização de análise de dados na indústria. Em seguida, a problemática industrial é apresentada e são trazidas as motivações para a utilização de técnicas de *Machine Learning* para a determinação de regiões operacionais ótimas. Por fim, são apresentados os objetivos do presente projeto, passando pelos objetivos gerais e específicos.

1.1 Contexto

O termo revolução industrial é atribuído ao desenvolvimento de novas tecnologias que provocam mudanças rápidas e profundas nos sistemas econômicos e na estrutura social [26]. Atualmente, o mundo passa pela chamada quarta revolução industrial, também conhecida como Indústria 4.0 que é relacionada à aplicação de tecnologias digitais na indústria. Nesse aspecto, o uso de Inteligência Artificial, *Artificial Intelligence* (AI), é peça chave nessa mudança.

Na indústria, o uso de análise de dados tem potencial de melhorar a produtividade, a qualidade, a eficiência e a flexibilidade dos sistemas de produção. Assim, eles são a matéria-prima que entram em um fluxo até chegarem como informação aos tomadores de decisão. Esse fluxo de dados é representado na Figura 1.1



Figura 1.1: Fluxo de dados na Indústria 4.0. FONTE: Larrañaga et al. (2018) [13]

A etapa de transformação consiste na aplicação de técnicas de análise de dados para a transformação deles em conhecimento, o qual pode ser usado pelos tomadores de decisão. O estudo apresentado neste trabalho foca na transformação dos dados via modelos de *Machine Learning* para a obtenção de informações pertinentes para a indústria.

Em processos industriais automatizados, os equipamentos industriais são

munidos de uma série de instrumentos de medição, controle e sistemas de supervisão capazes de medir e controlar as condições operacionais de cada equipamento constantemente. Esses dados são coletados e armazenados em sistemas de gerenciamento de informações.

1.2 Motivação

Apesar dos grandes avanços da automatização dos sistemas industriais, os mesmos ainda estão sujeitos a falhas que podem levar a prejuízos financeiros, além de danos ao meio ambiente e à saúde humana. Nesse sentido, o gerenciamento de eventos anormais é uma área de grande interesse no controle de processos, sendo a Detecção e o Diagnóstico de Falhas (FDD - *Fault Detection and Diagnosis*) o principal elemento.

Devido à complexidade dos processos modernos e a necessidade de diagnósticos rápidos, a dependência de operadores humanos para lidar com esses eventos anormais não é adequada. Assim, a automatização de sistemas de FDD é uma importante linha de pesquisa atual.

Existem diversas técnicas de FDD que podem ser classificadas em: (i) métodos baseados em modelos e (ii) métodos baseados em dados. Métodos baseados em modelos são aqueles desenvolvidos com conhecimentos prévios fundamentais do processo, porém, devido à complexidade dos sistemas atuais, o desenvolvimento de modelos exatos é uma tarefa difícil. Métodos baseados em dados utilizam apenas informações obtidas dos dados do processo, sem necessidade de modelos ou conhecimento especializado [32]. Devido à complexidade de processos industriais, métodos baseados em dados são mais aplicáveis no gerenciamento de riscos [4]. Nesse aspecto, técnicas de *Machine Learning* podem ser utilizadas para a modelagem do sistema.

O estudo aqui apresentado relaciona-se à utilização de técnicas baseadas em dados para a detecção e diagnóstico da formação de produtos poliméricos *Off-Spec*, i.e., fora de especificação. Os dados utilizados nesse trabalho vêm da malha de controle de um determinado equipamento industrial. O produto final deve atender a certas exigências de qualidade, uma dessas exigências é a quantidade de imperfeições

presentes na resina. Reduzir ou eliminar a formação dessas imperfeições durante o processo industrial pode ser uma tarefa difícil devido à complexidade do problema.

Essa complexidade está diretamente relacionada a grande quantidade de variáveis medidas durante o processo. Devido a natureza multivariável do problema em questão, técnicas de *Machine Learning* de redução de dimensionalidade foram estudadas.

1.3 Objetivos

O objetivo geral do trabalho é utilizar técnicas de redução de dimensionalidade para detectar regiões operacionais que minimizem a formação de imperfeições em uma dada resina polimérica.

Os objetivos específicos são:

- Determinar as variáveis que se relacionam ao problema de imperfeições, evitando incluir no modelo variáveis que possam atrapalhar a resolução do problema ou a inclusão de variáveis não relacionadas ao problema;
- Realizar uma redução de dimensão do problema de modo que as principais informações continuem retidas nas variáveis resultantes. Com menor quantidade de variáveis, a explicabilidade do modelo será maior;
- Determinar regiões operacionais que minimizem a quantidade de imperfeições do produto final;
- Determinar as variáveis de processo com maior importância para a determinação das regiões ótimas;
- Comparar os modelos desenvolvidos.

2 Revisão Bibliográfica

Neste capítulo, é feito um levantamento da literatura científica para melhor compreensão do assunto abordado neste trabalho. Primeiramente, será apresentado o contexto do uso de *Machine Learning* na indústria, em seguida serão detalhadas as técnicas de gerenciamento de eventos anormais. Por fim, serão apresentados alguns métodos de pré-processamento de dados que são importantes para o estudo em questão.

2.1 *Machine Learning* na indústria

Nos dias atuais, a grande quantidade de dados dos processos que as indústrias possuem pode ser um fator chave para o aumento da produtividade, competitividade e crescimento de uma companhia. A mineração e a análise de dados são ferramentas capazes de extrair informações, reconhecer padrões de dados e realizar previsões. *Machine Learning* é um conjunto de técnicas que permite a realização dessas análises.

Machine Learning é usado para se chegar a conclusões úteis a partir de dados. Segundo Larrañaga et al. (2018) [13], existem três tipos principais de conclusões: (i) *clusterização*, com o objetivo de encontrar *clusters* de *inputs* semelhantes; (ii) previsão, prevendo o *output* para futuros *inputs*; e (iii) descoberta de associação, procurando relações (probabilísticas) entre *inputs* e *outputs*.

O Processo Padrão Inter-Indústrias para Mineração de Dados (CRISP-DM) é um modelo que descreve o processo comumente usado para a transformação de dados em modelos de *Machine Learning* na indústria. O CRISP-DM é composto por seis etapas:

1. **Compreensão de Negócio:** Primeira etapa que consiste na compreensão do objetivo do projeto do ponto de vista do negócio;
2. **Compreensão dos Dados:** Nesta fase, os dados do problema são coletados e há uma familiarização com esses para se descobrir os primeiros

insights;

3. **Preparação dos Dados:** Etapa de pré-processamento e construção do conjunto de dados final a partir do conjunto inicial;
4. **Modelagem:** Nessa etapa, técnicas de *Machine Learning* são aplicadas;
5. **Avaliação:** O modelo desenvolvido é avaliado segundo sua performance.
6. **Implementação:** Se a performance do modelo é igual ou acima da esperada, o modelo desenvolvido é implementado na indústria.

A Figura 2.1 mostra as seis etapas descritas no processo CRISP-DM.

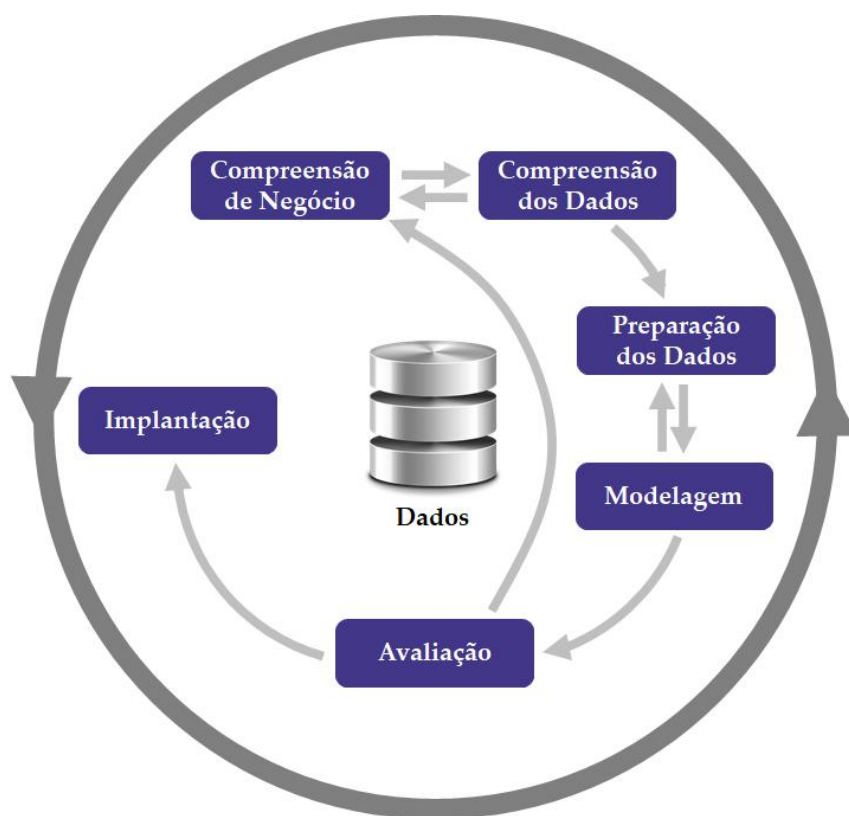


Figura 2.1: As seis etapas de um processo CRISP-DM que transforma um conjunto de dados em conhecimento para uma empresa. FONTE: Shearer (2000)[27]

Com a crescente necessidade de processos mais seguros e que permitam a fabricação de produtos de alta qualidade, a detecção e diagnóstico de falhas é uma área em ascensão na indústria e faz uso de técnicas de *Machine Learning*. Na próxima seção, esse tópico será abordado em mais detalhes.

2.2 Detecção e Diagnóstico de Falhas

O gerenciamento de eventos anormais é essencial para se manter condições operacionais favoráveis e se prever riscos de processos químicos [32]. Neste aspecto, métodos de detecção e diagnóstico de falhas (FDD) têm como objetivo principal detectar falhas e diagnosticar suas causas para permitir a correção antes que danos maiores sejam causados. Isermann (2006) [9] define uma falha como "um desvio não permitido de pelo menos uma propriedade característica (*feature*) do sistema da condição operacional normal e aceitável". De acordo com Ding (2010) [6], a FDD segue os três procedimentos abaixo:

- **Detecção de falha:** Detecta que há algo de errado na unidade funcional analisada do processo;
- **Isolamento de falha:** Detecta a localização e a classificação da falha;
- **Identificação de falha:** Determina a magnitude da falha.

A ocorrência de eventos anormais pode ser ocasionada por falhas em componentes do equipamento de processo, falhas do sistema de controle, recursos que estão em falta, alteração devido à não conformidade do produto, dentre outras causas [16]. Após a detecção de uma anormalidade, a causa pode então ser localizada e identificada para a tomada de decisões.

Na literatura científica existem diversas terminologias para classificar os métodos de FDD. Neste trabalho iremos considerar a terminologia adotada por Venkatasubramanian et al. (2003) [30, 31, 32] que está sendo representada na Figura 2.2.

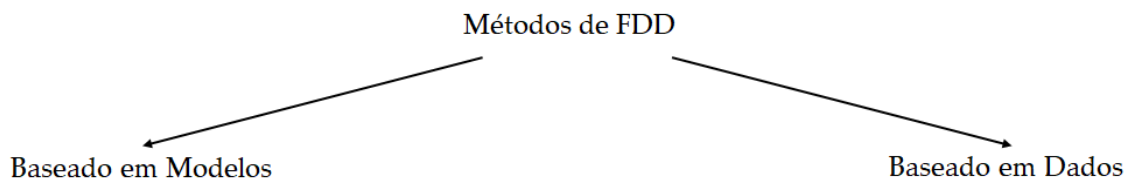


Figura 2.2: Classificação de Métodos de FDD

O método a ser usado para determinada análise é escolhido baseado no sistema em questão e nas informações disponíveis sobre ele. Apesar disso, Venkata-

subramanian et al. (2003) [32] trazem características comuns desejáveis para todos os métodos de detecção e diagnóstico de falhas, algumas delas são:

- Rápida detecção e diagnóstico;
- Capacidade de distinguir entre diferentes falhas;
- Robustez;
- Identificabilidade de novidade: essa característica relaciona-se a capacidade de reconhecer a ocorrência de novas falhas e não as classificar erroneamente como um problema conhecido ou como operação normal.

O estudo presente usou conceitos de FDD para determinar condições operacionais que levam à formação de uma menor quantidade de imperfeições num produto polimérico. Para isso, um levantamento dos principais métodos de FDD foi realizado e será apresentado de forma sucinta nas seções subsequentes.

2.2.1 Métodos Baseados em Modelos

Os métodos baseados em modelo podem ser divididos em métodos quantitativos e qualitativos. Essencialmente a técnica quantitativa utiliza um modelo matemático de redundância analítica. Por outro lado, métodos qualitativos são geralmente desenvolvidos com base em alguns conhecimentos fundamentais da física e da química do processo [32, 30]. Nesta seção, serão apresentados mais a fundo os métodos por redundância analítica.

2.2.1.1 Redundância Analítica

Os métodos quantitativos baseados em modelo apresentam uma redundância analítica onde é comparado o *output* do processo monitorado com o *output* de um modelo matemático analítico. Essa técnica envolve duas etapas: (i) geração de resíduo e (ii) avaliação do resíduo. Esse resíduo corresponde à inconsistência em relação ao modelo do processo quando comparado ao modelo esperado [32, 18].

Um esquema desse método de FDD é representado na Figura 2.3. Com base nos sinais de entrada medidos \mathbf{U} e nos sinais de saída \mathbf{Y} , os métodos de detecção geram valores residuais \mathbf{r} , estimativas de parâmetros Θ ou estimativas de estado \hat{x} , que são chamados de *features*. Alterações dos *features* em relação aos seus valores nominais são detectadas, levando a sintomas analíticos \mathbf{s} .

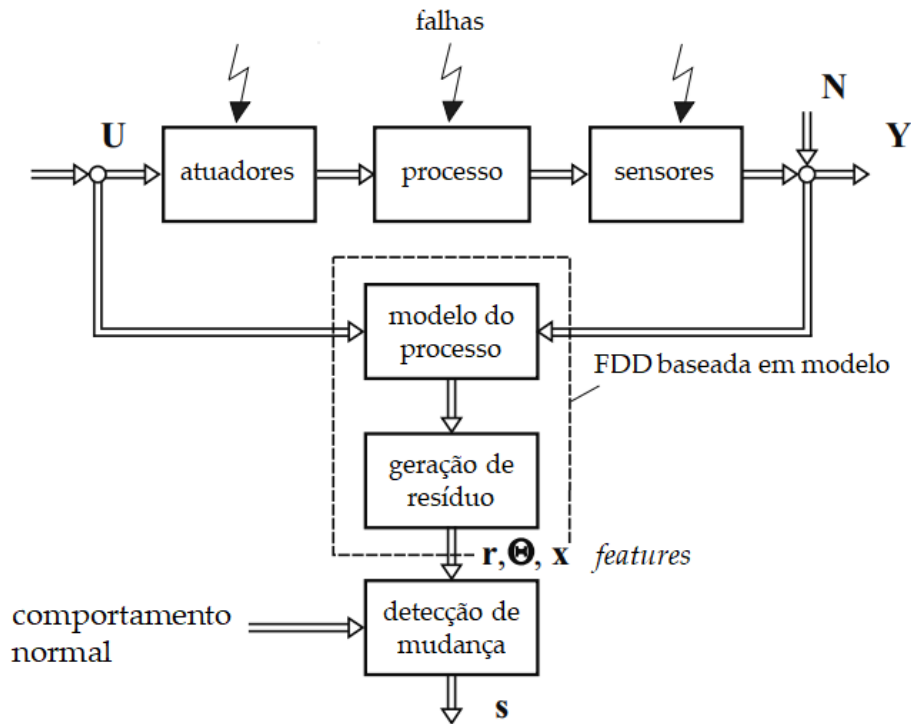


Figura 2.3: Esquema da detecção de falhas baseada em modelo. FONTE: Isermann (2006) [9]

Algumas das técnicas mais comuns de métodos baseados em modelos de redundância analítica são apresentadas na sequência.

i) Equação de Paridade

Nesse método, utilizando os sinais de entrada do modelo \mathbf{U} é possível estimar os valores de saída \mathbf{Y}_m do processo através do modelo do processo. Esses valores são então comparados ao *output* do processo \mathbf{Y} , gerando um resíduo \mathbf{r} . Um esquema dessa abordagem é representado na Figura 2.4. O processo é descrito pela função de transferência \mathbf{G}_p e o modelo matemático analítico é descrito pela função \mathbf{G}_m , dessa forma, o resíduo é calculado através da Equação 2.1 [8].

$$r(s) = (G_p - G_m) U(s) \quad (2.1)$$

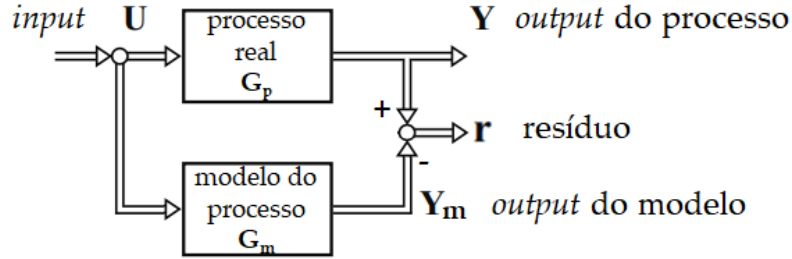


Figura 2.4: Esquema da detecção de falhas por equações de paridade. FONTE: Miljković (2011) [18]

ii) Estimativas de parâmetros

Na maior parte dos casos práticos, parâmetros do processo são parcialmente ou completamente desconhecidos. No caso em que a estrutura do modelo matemático do processo é conhecida, esses parâmetros Θ podem ser estimados através apenas dos sinais de entrada U e de saída Y conhecidos:

$$Y(s) = f(U(s), \Theta) \quad (2.2)$$

Esses parâmetros estimados Θ são relacionados à parâmetros físicos do sistema ϕ (como por exemplo atrito, massa, resistência, capacitância) através de $\Theta = g(\phi)$. Então, parâmetros $\hat{\phi}$ são estimados através de $\hat{\phi} = g^{-1}(\Theta)$, dessa forma o desvio $\Delta\phi$ é computado e, usando métodos de reconhecimento de padrões, é possível relacionar as mudanças $\Delta\phi$ às falhas do processo [32].

iii) Filtro de Kalman

O filtro de Kalman é um método que calcula o resíduo de variáveis de estado de um sistema dinâmico linear a partir de sinais de entrada U e saída Y . A teoria geral desse método é descrita na sequência.

Um sistema estocástico de dimensão finita linear é definido pelo seguinte modelo de espaço de estado [5]:

$$\mathbf{X}(k+1) = \mathbf{A}\mathbf{X}(k) + \mathbf{B}\mathbf{U}(k) + w(k) \quad (2.3)$$

$$\mathbf{Y}(k) = \mathbf{C}\mathbf{X}(k) + v(k) \quad (2.4)$$

Onde $\mathbf{X}(k)$, $\mathbf{U}(k)$ e $\mathbf{Y}(k)$ representam o estado, o *input* e o *output*, respectivamente. \mathbf{A} , \mathbf{B} e \mathbf{C} são matrizes de tamanhos apropriados. $w(k)$ e $v(k)$ são os ruídos gaussianos brancos do processo e da medida, respectivamente.

O esquema recursivo para estimativa de estado ideal é [8]:

$$\hat{\mathbf{X}}(0|0) = \bar{\mathbf{X}} \quad (2.5)$$

$$\hat{\mathbf{X}}(k|k-1) = \mathbf{A}\hat{\mathbf{X}}(k-1|k-1) + \mathbf{B}\mathbf{U}(k-1) \quad (2.6)$$

$$\hat{\mathbf{X}}(k|k) = \hat{\mathbf{X}}(k|k-1) + \mathbf{K}\mathbf{e}(k) \quad (2.7)$$

$\hat{\mathbf{X}}(i|j)$ representa a estimativa de $\hat{\mathbf{X}}(i)$ dadas as medidas $Y(1), \dots, Y(j)$. O erro da estimação $\mathbf{e}(k)$ é dado por $\mathbf{Y}(k) - \mathbf{C}\hat{\mathbf{X}}(k|k-1)$. Finalmente, \mathbf{K} é o ganho do filtro de Kalman que é calculado a partir das matrizes de covariância dos ruídos e de erro de estado. Mais detalhes podem ser encontrados em Ding (2008) [5] e Venkatasubramanian et al. (2003) [32].

O filtro de Kalman é aplicável apenas para sistemas lineares, entretanto, boa parte dos sistemas são não lineares. Com o objetivo de utilizar o filtro de Kalman em sistemas não lineares, foi desenvolvido o filtro de Kalman estendido. Nesse método, a expansão da série de Taylor de primeira ordem é utilizada para aproximar sistemas não lineares. Entretanto, se essa técnica é aplicada em sistemas que têm fatores de ordem alta, esse método aumentará o erro do sistema [15]. Outro método utilizando o filtro de Kalman que foi desenvolvido por Julier e Uhlmann [12] é o Filtro de Kalman *Unscented* que se baseia na transformação *Unscented*.

2.2.2 Métodos Baseados em Dados

Em métodos baseados em dados, é necessário ter a disposição uma grande quantidade de dados históricos do processo, porém, ao contrário dos métodos baseados em modelos, não é necessário conhecimentos a priori sobre o processo. Ven-

katasubramanian et al. (2003) divide esses métodos também em quantitativos e qualitativos, relacionando-os ao tipo de extração de dados [31].

Dentre os métodos qualitativos mais conhecidos estão o sistema especialista e a lógica Fuzzy. O sistema especialista, entretanto, apresenta aplicação limitada na indústria química [4]. Métodos quantitativos avaliam problemas de diagnósticos de falhas como um problema de reconhecimento de padrões, assim, os pontos de dados são classificados em classes pré-determinadas [31].

Em geral, métodos quantitativos são relacionados a métodos estatísticos uni ou multivariados. Na literatura, encontram-se diversos métodos de análise estatística multivariada como PCA, PLS, ICA e FDA.

Em problemas de processos químicos, geralmente, tem-se problemas multivariados, i.e. problemas com várias *TAGs* (etiquetas de identificação) de processo. Métodos de análise estatística multivariada são ferramentas poderosas que permitem a redução de dimensionalidade do problema, retendo as informações principais e facilitando a análise [31]. Em geral, essa abordagem usa a estatística T^2 de Hotelling e a estatística Q , essa última também conhecida como erro de predição ao quadrado (SPE) [21].

2.2.2.1 Estatística T^2 de Hotelling

A estatística T^2 de Hotelling é frequentemente usada em gráficos de controle. Essa estatística se baseia na generalização da estatística t de Student para o caso multivariado. Considerando \mathbf{x}_i como o vetor linha de tamanho m que representa as medições de todos os sensores na amostra i . Seja \mathbf{t}_{si} o vetor linha de tamanho r , que é a projeção do experimento \mathbf{x}_i no novo espaço. Esses dois vetores estão relacionados da seguinte forma:

$$\mathbf{t}_{si} = \mathbf{x}_i \mathbf{P} \quad (2.8)$$

Onde \mathbf{P} é a matriz de transformação linear $m \times r$, sendo n o número de amostras. A estatística T^2 da i -ésima amostra é definida como:

$$T_i^2 = \sum_{j=1}^r \frac{t_{sj}^2}{\lambda_j} = \mathbf{t}_{si} \Lambda^{-1} \mathbf{t}_{si}^T = \mathbf{x}_i \mathbf{P} \Lambda^{-1} \mathbf{P}^T \mathbf{x}_i^T \quad (2.9)$$

Onde Λ é a matriz diagonal $\Lambda = \text{diag}(\lambda_1, \lambda_2, \dots, \lambda_m)$ e λ_j são os autovalores da matriz de covariância cujos autovetores compõe a matriz de transformação linear \mathbf{P} . Uma falha é detectada quando o valor de T^2 excede um limite específico, onde o desvio do plano normal pode ser observado. A estatística T^2 é uma medida da variação de cada amostra dentro do modelo de redução de dimensão. [20]

2.2.2.2 Estatística Q

A estatística Q também pode ser usada para a detecção de falhas. Essa estatística representa a distância residual, ou quadrada, de uma nova amostra do plano do modelo, que mede a soma total da variação no espaço residual [21].

A estatística Q denota a mudança dos eventos que não são explicados pelo modelo de componentes principais. É uma medida da diferença entre uma amostra e sua projeção no modelo. A estatística Q_i do vetor de amostra \mathbf{x}_i é definida da seguinte forma:

$$Q_i = \tilde{\mathbf{x}}_i \tilde{\mathbf{x}}_i^T = \mathbf{x}_i (\mathbf{I} - \mathbf{P} \mathbf{P}^T) \mathbf{x}_i^T \quad (2.10)$$

Onde $\tilde{\mathbf{x}}_i$ é a projeção do vetor \mathbf{x}_i no espaço residual e \mathbf{I} é a matriz identidade $m \times m$. Normalmente, a estatística Q é mais sensível do que a estatística T^2 , isso porque T^2 tem grande variância e, portanto, requer uma grande mudança nas características do sistema para a detecção de uma falha. [20]

Informações sobre os eventos anormais também podem ser obtidas diretamente do gráfico das componentes principais relevantes. Quando há uma mudança no sistema, os valores das componentes principais dos novos eventos serão diferentes dos valores anteriores e a mudança será detectada. [20]

2.2.2.3 Análise de Componentes Principais - PCA

A análise de componentes principais é uma técnica de redução de dimensões que decompõe a matriz de covariância dos dados de processo nas direções de variância máxima dos dados, i.e., ela transforma uma distribuição multivariável correlacionada em combinações lineares ortogonais das variáveis iniciais. A PCA é uma importante ferramenta de monitoramento de processos graças a simplicidade e eficiência em processar grandes quantidades de dados [6].

Após realizar decomposição usando PCA é possível utilizar as estatísticas T^2 de Hotelling e SPE nos novos vetores obtidos para determinar quando uma falha ocorre ou não, como é possível observar no trabalho de Yin et al. (2012) [33] que apresenta uma comparação de alguns métodos baseados em dados. Nesse trabalho, os autores comparam os métodos no processo *Tennessee Eastman*.

No estudo de Tao et al. (2013) [29], os autores realizam um estudo de detecção de falhas em um processo de reator de batelada sequencial para o tratamento de água residual. Nesse trabalho, observa-se a detecção de falhas via análises estatísticas de T^2 de Hotelling e SPE, e também é realizada uma análise dos gráficos das componentes de PCA.

2.2.2.4 Mínimos Quadrados Parciais - PLS

Os mínimos quadrados parciais é uma outra técnica de redução de dimensão linear muito usada na detecção de falhas em processos industriais. PCA é um modelo baseado no *input*, PLS é um modelo baseado no *input-output* e pode ser usado para detectar falhas tanto no processo quanto nas variáveis [1]. O método de PLS reduz a dimensão do problema enquanto maximiza a covariância entre a matriz preditora (*input*) \mathbf{U} e a matriz predita (*output*) \mathbf{Y} para cada componente do espaço reduzido.

2.2.2.5 Análise de Componentes Independentes - ICA

A ICA decompõe os dados em combinações lineares de componentes estatisticamente independentes, chamadas de ICs. O objetivo desse método é minimizar a

dependência estatística entre os vetores de base. Enquanto PCA comprime informações, ICA separa informações independentes dos dados de entrada. ICA é indicada para casos de tratamento de sinais em casos de *Blind Source Separation* (BSS) que consistem em problemas nos quais sinais misturados são conhecidos, porém se desconhece os sinais de origem e o processo de mistura [14]. Em geral, ICA utiliza estatística de T^2 e SPE para FDD [21].

2.2.2.6 Análise de Discriminante de Fisher - FDA

A análise de discriminante de Fisher, assim como PCA, é uma técnica linear de redução de dimensão. A diferença entre as duas técnicas é que FDA leva em consideração as informações sobre a classificação das amostras [24]. PCA é uma técnica que encontra direções que maximizam a variância, já FDA busca encontrar subespaços que maximizam a separação de classes.

2.2.3 Métodos de clusterização

A detecção de falhas baseadas em estatísticas está normalmente associada a detecção de falhas de sensores e pode não apresentar resultados desejados quando se trata de análise de qualidade de produto. Uma das alternativas encontradas é a utilização de clusterização para a identificação de classes.

A clusterização é um dos métodos de *Machine Learning* mais utilizados. Métodos de clusterização permitem agrupar dados em *clusters* semelhantes. A clusterização é usada para extrair padrões para análise de dados, classificação de padrões, agrupamento, tomada de decisão, aprendizado de máquina, mineração de dados, recuperação de documentos, segmentação de imagens e outras aplicações. [10]

Vários métodos de clusterização são encontrados, dentre eles um dos mais conhecidos por sua simplicidade é *KMeans*. A performance dos métodos de clusterização está diretamente relacionada à complexidade do problema e dos dados. O uso de métodos combinados mostra-se como uma alternativa para lidar com problemas mais complexos.

Métodos combinados de clusterização utilizando PCA e *KMeans* são frequentemente encontrados na literatura científica para a identificação de padrões e previsões. PCA é utilizado para redução de dimensões de conjuntos de dados com grande número de variáveis e *KMeans* é usado para detecção de *clusters* similares, sendo possível, assim, a visualização dos dados.

Como exemplo da aplicação dessas técnicas, PCA seguido de *KMeans* foi utilizado para a previsão de câncer de mama no trabalho de Jamal et al. (2018) [11]. Também é possível observar aplicações semelhantes no estudo de Tao et al. (2013) [29], onde os autores apresentam a visualização de *clusters* de falhas em gráficos das componentes principais de PCA para um processo de um reator de batelada sequencial.

Neste trabalho, será utilizada a redução de dimensões via PCA e posteriormente a clusterização via *KMeans* para a identificação de regiões que minimizam a formação de imperfeições em um processo industrial de produção de polímero. Em seguida, uma outra abordagem proposta com a redução de dimensões via FDA e clusterização via *KMeans* é apresentada. Por fim, uma comparação de ambos os métodos é feita.

2.3 Seleção de Variáveis

Processos químicos modernos apresentam uma grande quantidade de variáveis de processos. Entretanto, nem todas as variáveis medidas são igualmente importantes para desenvolver bons modelos de monitoramento de eventos anormais. Normalmente, apenas um conjunto dessas variáveis é necessário para modelar o problema. A seleção de variáveis torna-se, assim, uma pré-etapa importante para a FDD. Essa seleção permite redução de custo computacional, aumento da velocidade e da precisão e melhora da compreensão dos modelos ao reduzir a dimensionalidade dos dados e eliminar variáveis redundantes e irrelevantes. [7]

Na literatura é possível encontrar diversos algoritmos para seleção de variáveis. Esses algoritmos podem ser divididos em duas classes: (i) Métodos de filtro e (ii) Métodos de envelope. Métodos de filtro utilizam propriedades intrínsecas dos

dados para seleção de variáveis. Já métodos de envelope utilizam o algoritmo que será aplicado nos dados na sequência como critério de seleção de variáveis. [28]

Métodos de filtro são apreciados pela sua simplicidade e generalidade e, em geral, são baseados em critérios estatísticos. Por outro lado, os métodos por envelope exigem uma capacidade computacional maior e são específicos para o problema em análise, em geral, sendo utilizados algoritmos genéticos. [7, 28]

Neste trabalho, como uma etapa de pré-processamento, serão utilizados métodos de filtro para a seleção de variáveis.

2.4 Importância das Variáveis

Em *Machine Learning*, interpretabilidade é a capacidade de explicar em termos compreensíveis para um ser humano. Quando existem investimentos, as empresas, em geral, preferem modelos que sejam explicáveis em vez de modelos que possam fornecer uma precisão relativamente melhor [22]. Dessa forma, é necessário adicionar ao modelo um novo parâmetro, a explicabilidade.

Uma das formas de aumentar a explicabilidade de um modelo é através da análise da importância das variáveis. Ao analisar esse parâmetro, é possível observar qual variável mais impacta na tomada de decisão do modelo desenvolvido.

Devido a complexidade crescente dos modelos de *Machine Learning* desenvolvidos, há uma necessidade de métodos mais robustos de determinação de importância de variáveis. Estudos utilizando teoria dos jogos para esse fim apresentam resultados interessantes que tornam os modelos desenvolvidos mais interpretáveis [17]. A teoria dos jogos é uma teoria matemática criada para se modelar fenômenos que podem ser observados quando dois ou mais “agentes de decisão” interagem entre si [25].

Valor *Shapley* é um conceito originalmente desenvolvido para estimar a importância de um jogador individual em uma equipe. No contexto de previsões de modelo, os valores *Shapley* também podem ser compreendidos como uma alocação justa ou razoável da importância da variável dado um determinado modelo de saída

[23].

Como forma de aumentar a explicabilidade dos modelos desenvolvidos neste projeto, a análise de importância das variáveis via valor *Shapley* será apresentada.

3 Metodologia

Após todo levantamento teórico mostrado no Capítulo 2, neste capítulo será apresentada e detalhada a Metodologia. Inicialmente será descrito de forma sucinta o sistema que foi utilizado para a realização do estudo de caso. Em seguida, o processamento realizado nos dados é apresentado. Por fim, a atenção será voltada para as estruturas de detecção de regiões operacionais que minimizam a formação de imperfeições no produto final.

3.1 Sistema

O estudo foi realizado em cima dos dados obtidos de um equipamento industrial que tem como fim a produção de resinas poliméricas. Esse equipamento apresenta duas zonas que são munidas de diversos instrumentos de medição e controle, essas variáveis de processo (*TAGs*) são armazenadas em um banco de dados. As *TAGs* desse processo são apresentadas na Tabela 3.1.

Tabela 3.1: *TAGs* do processo

<i>TAG</i>	Descrição
A0001	O2 no equipamento
L0101	Nível 1
L0201	Nível 2
F0001	Vazão de alimentação 0
F0101	Vazão de alimentação 1
F0201	Vazão de alimentação 2
F0301	Vazão de alimentação 3
F0301R	Razão das Vazões 3/1
F0401	Vazão de alimentação 4
F0401R	Razão das Vazões 4/1
F0501	Vazão de alimentação 5
F0501R	Razão das Vazões 5/1
T1001	Temperatura 0 da Zona 1
T1101	Temperatura 1 da Zona 1
T1201	Temperatura 2 da Zona 1
T1301	Temperatura 3 da Zona 1
T1401	Temperatura 4 da Zona 1

T1501	Temperatura 5 da Zona 1
T1601	Temperatura 6 da Zona 1
P1001	Pressão 0 da Zona 1
P1101	Pressão 1 da Zona 1
P1201	Pressão 2 da Zona 1
P1301	Pressão 3 da Zona 1
P1401	Pressão 4 da Zona 1
J1001	Potência
J1101	Potência Específica
Z1001	Abertura
Z1001SP	Set-point da Abertura
Z1001E	Erro da Abertura
T2001	Temperatura 0 da Zona 2
T2101	Temperatura 1 da Zona 2
T2201	Temperatura 2 da Zona 2
T2301	Temperatura 3 da Zona 2
T2401	Temperatura 4 da Zona 2
T2501	Temperatura 5 da Zona 2
T2601	Temperatura 6 da Zona 2
P2001	Pressão 0 da Zona 2
P2001SP	Set-point da Pressão 0 da Zona 2
P2001O	Out-put da Pressão 0 da Zona 2
P2101	Pressão 1 da Zona 2
P2201	Pressão 2 da Zona 2
P2301	Pressão 3 da Zona 2
P2401	Pressão 4 da Zona 2
P2401SP	Set-point da Pressão 4 da Zona 2
P2401O	Out-put da Pressão 4 da Zona 2
P2501	Pressão 5 da Zona 2
I2001	Corrente da Zona 2
C2001	Eficiência da Zona 2
D2001	Torque da Zona 2
S2001	Rotação da Zona 2

Para a produção de resinas poliméricas, é necessário realizar o seu controle de qualidade. Um dos parâmetros de qualidade medido no processo em questão é a quantidade de imperfeições do produto ao fim do processo. Essa quantidade é medida e seus valores são armazenados no banco de dados sob a TAG X3001. Duas outras TAGs categóricas foram criadas baseadas no valor da TAG contínua X3001.

Essas *TAGs* de qualidade são exibidas na tabela abaixo.

Tabela 3.3: *TAGs* de qualidade

<i>TAG</i>	Descrição
X3001	Quantidade de Imperfeições no produto final
CAT-SPEC	Categoria do produto em relação a imperfeições em On e Off-Spec
CAT-LEVELS	Categoria do produto em relação a imperfeições em Ótimo, Aceitável e Ruim

3.2 Recursos Utilizados

No tratamento dos dados e no desenvolvimento dos modelos foram utilizadas as seguintes ferramentas e linguagens:

- **Python:** É uma linguagem de alto nível orientada a objeto que tem grande popularidade devido a sua estrutura dinâmica e sintaxe clara;
- **NumPy:** Biblioteca do Python de análise numérica e funções matemáticas que opera com vetores e matrizes;
- **Pandas:** Biblioteca do Python usada para manipulação e análise de dados em forma de estruturas chamadas *DataFrames*;
- **Scikit-Learn:** Biblioteca da linguagem Python que apresenta um grande conjunto de ferramentas que auxiliam processos de *Machine Learning*;
- **Matplotlib:** : É uma biblioteca para construir gráficos da linguagem python de rápida implementação;
- **SHAP::** (SHapley Additive exPlanations) é uma abordagem que utiliza teoria dos jogos para explicar a saída de qualquer modelo de *Machine Learning*;
- **Anaconda:** : É uma plataforma de distribuição de linguagens Python e R que visa simplificar o gerenciamento e implementação dos pacotes;
- **Jupyter Notebook::** É um ambiente web de código aberto que permite criar e compartilhar documentos que contêm código ativo, equações, visualizações e texto narrativo.

A tabela Tabela 3.4 explicita quais foram as versões de cada uma das ferramentas utilizadas.

Tabela 3.4: Tabela com as versões de cada uma das ferramentas

Ferramenta	Versão
Python	3.7.4
Numpy	1.19.2
Pandas	1.1.3
Scikit-Learn	0.23.2
Matplotlib	3.3.2
SHAP	0.39.0

3.3 Processamento dos dados

3.3.1 Limpeza dos dados

Para se assegurar modelos melhores, é necessário fazer uma limpeza dos dados a priori. Os dados armazenados do processo podem apresentar qualidade *Good* ou *Bad*. A qualidade dos dados é a medida de quão confiáveis são os dados de uma determinada *TAG*. Se a qualidade de uma *TAG* não for *Good*, o valor geralmente não deve ser confiável. Há uma grande variedade de causas para dados *Bad*, como, por exemplo, desconexões de rede, falhas de software e configurações de *TAGs* inválidas. Dados *Bad* não são confiáveis, pois o valor armazenado pode não corresponder ao real valor da variável num determinado instante. Logo, inicialmente, removeu-se todos os valores com status *Bad*.

Uma segunda etapa de limpeza dos dados consistiu em estabelecer limites máximos e mínimos de cada variável para a remoção de possíveis *outliers*. Esses limites foram estabelecidos juntamente com a equipe de processos da planta. Por último, verificou-se quais pontos de medição de quantidade de imperfeições correspondiam a medições ruins de laboratório, sendo esses pontos removidos do conjunto de dados inicial.

3.3.2 Redimensionamento de variáveis

Como em processos industriais as variáveis medidas apresentam magnitudes distintas, existe a necessidade de se realizar um redimensionamento dessas variáveis na etapa de processamento para permitir a aplicação de modelos de análise de dados. Existem diferentes formas de redimensionar vetores, neste trabalho foi utilizado o redimensionamento via padronização, já que os modelos de PCA e FDA foram construídos baseados em vetores redimensionados dessa forma.

O redimensionamento, via padronização, de cada vetor de amostras \mathbf{x} da variável de processo x foi feito da seguinte forma:

$$\mathbf{z} = \frac{\mathbf{x} - \mu}{\sigma} \quad (3.1)$$

Onde μ e σ correspondem a média e ao desvio padrão do vetor de amostras da variável \mathbf{x} e \mathbf{z} é o vetor padronizado. Inicialmente, realizou-se uma padronização de todas as variáveis de processo e as análises de dados seguintes foram feitas com esses dados padronizados. Esse recurso foi aplicado utilizando a função *StandardScaler* do pacote *preprocessing* da biblioteca *scikit-learn*.

3.3.3 Seleção de Variáveis

O equipamento industrial em análise apresenta cinquenta variáveis de monitoramento (*TAGs*). Devido à grande quantidade de variáveis de processo, é necessário realizar uma seleção das variáveis que são relevantes para o problema. Como já discutido na Seção 2.3, existem métodos de filtro e métodos de envelope. Em razão da simplicidade e do caráter generalista dos métodos de filtro, optou-se por usar um desses métodos para realizar essa seleção.

Para isso, analisou-se a correlação entre as variáveis do processo e a variável objetivo que mede a quantidade de imperfeições do produto. O coeficiente de correlação r mede a associação entre duas variáveis:

- $r = 0$, não existe correlação;
- $r = +1$, as variáveis estão perfeitamente correlacionadas positivamente;

- $r = -1$, as variáveis estão perfeitamente correlacionadas negativamente.

Dessa forma, se o valor absoluto do coeficiente de correlação entre duas variáveis é próximo a 1, podemos dizer que essas duas variáveis estão fortemente correlacionadas. A correlação de Pearson é uma das mais conhecidas e mede a correlação linear entre duas variáveis. Neste trabalho, usou-se a correlação de Spearman, que avalia a relação monotônica entre duas variáveis contínuas ou ordinais. Nesse caso, as variáveis tendem a mudar juntas, mas não necessariamente a uma taxa constante [19].

A correlação de Spearman, também conhecida como correlação de *rank*, calcula a correlação entre o *rank* de duas variáveis (\mathbf{x} e \mathbf{y}). Ao invés dos valores de \mathbf{x} e \mathbf{y} , a classificação (*rank*) de cada variável é usada. *Rank* é o valor que cada variável possui quando ordenamos numericamente seus valores. A amostra com menor valor terá *rank* um, com segundo menor valor terá *rank* dois e assim por diante. Portanto, todas as variáveis terão novos valores, os chamados *ranks*. Caso não existam valores repetidos, o coeficiente de correlação é calculado pela equação abaixo:

$$r_s = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)} \quad (3.2)$$

Na Equação 3.2, d corresponde a diferença entre os ranks de duas amostras e n corresponde ao número de amostras. Caso existam valores repetidos, o coeficiente de correlação de Spearman é calculado da seguinte forma:

$$r_s = \frac{\sum R(x_i)R(y_i) - \frac{(\sum R(x_i))(\sum R(y_i))}{n}}{\sqrt{\left(\sum R(x_i)^2 - \frac{(\sum R(x_i))^2}{n}\right) \left(\sum R(y_i)^2 - \frac{(\sum R(y_i))^2}{n}\right)}} \quad (3.3)$$

Na Equação 3.3, $R(x_i)$ e $R(y_i)$ correspondem aos *ranks* das variáveis x_i e y_i , respectivamente.

Baseado nos coeficientes de correlação de Spearman entre as variáveis de processo e a variável de qualidade, fez-se uma pré seleção das variáveis de processo a serem usadas no modelo. Em seguida, analisou-se a correlação dessas variáveis de processo pré-selecionadas para eliminação de variáveis redundantes. Essa correlação foi feita utilizando a função *corr* da biblioteca Panda, especificando-se o método de Spearman.

3.3.4 Separação de Dados em Treino e Teste

Para se identificar possíveis problemas de *overfitting* do modelo, é necessário se separar os dados em treino e teste. *Overfitting* significa que o modelo que foi treinado se ajustou muito bem ao conjunto de dados de treino sendo, provavelmente, um modelo pouco generalizável que terá resultados muito bons para os dados de treino e resultados ruins para os dados de teste [2].

Para se observar se ocorreu *overfitting* dos modelos desenvolvidos nesse projeto, um terço dos dados iniciais foram separados como dados de teste e o restante dos dados foram usados para o treinamento. Essa separação de dados de treino e teste foi feita utilizando a função *train_test_split* do pacote *model_selection* da biblioteca *scikit-learn*. Os modelos foram então ajustados para os dados de treino e analisou-se o desempenho dos mesmos nos dados de teste.

3.4 Modelos de Identificação de Regiões que Minimizam Imperfeições

No Capítulo 2, foi apresentado um levantamento da literatura a respeito da detecção e diagnóstico de falhas em sistemas industriais. Neste trabalho, foram usados conceitos de FDD para a detecção de condições operacionais que conduzem à minimização da formação de imperfeições na produção de resinas poliméricas em um equipamento industrial. Para a redução de dimensionalidade, escolheu-se, primeiramente, utilizar a técnica de PCA, por ser a mais bem conhecida e documentada na literatura. Em seguida, a título de comparação, utilizou-se FDA porque esse método tem como objetivo a separação de classes, que é o mesmo problema abordado neste trabalho.

A Figura 3.1 representa a estrutura da abordagem utilizada neste projeto para a determinação de regiões operacionais que minimizam a quantidade de imperfeições em resinas poliméricas em um equipamento industrial.

Após a seleção das variáveis de processo, os dados não redimensionados são separados em conjunto de treino e conjunto de teste. Para evitar *Data Leakage*, o redimensionamento dos dados é ajustado apenas no conjunto de treino. *Data*

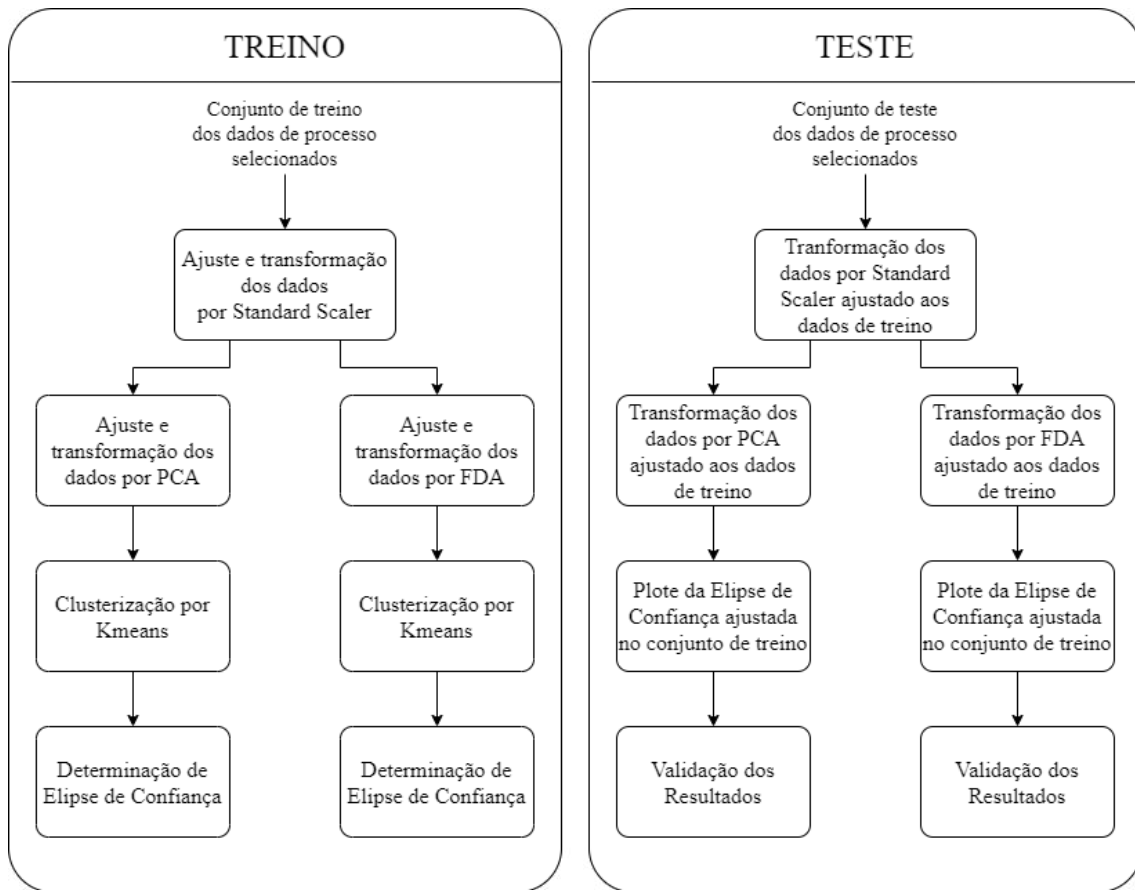


Figura 3.1: Estrutura do método de detecção de regiões operacionais que minimizam a quantidade de imperfeições utilizada nesse projeto

Leakage é um termo utilizado para quando informações fora do conjunto de treino são usadas para se criar o modelo [3]. Após o ajuste, a transformação é feita usando a Equação 3.1, onde μ e σ representam, respectivamente, a média e o desvio padrão das variáveis do conjunto de treino.

Analisando o fluxo de dados de treino exibido na Figura 3.1, vemos que o conjunto de treino segue por duas ramificações: PCA e FDA. Na ramificação de PCA, é ajustado um modelo de PCA baseado nos dados de treino redimensionados. Esses dados são transformados e, em seguida, um gráfico da componente 1 versus a componente 2 é feito. Os pontos de dados presentes nesse gráfico são clusterizados via *Kmeans* e, finalmente, uma elipse de confiança é gerada em torno do *cluster* com menor média de quantidade de imperfeições, assim determinando uma região operacional que minimiza a quantidade de imperfeições no produto final. A ramificação de FDA segue um fluxo semelhante, porém, ao invés de ser ajustado um modelo de

PCA, um modelo de FDA é ajustado aos dados de treino redimensionados.

Em relação ao fluxo de dados de teste, o conjunto de dados é redimensionado baseado no modelo ajustado ao conjunto de treino. Os dados redimensionados seguem então duas ramificações, assim como o conjunto de treino: PCA e FDA. Na ramificação de PCA, os dados são transformados baseado no modelo ajustado ao conjunto de treino, em seguida, um gráfico da componente 1 versus a componente 2 é gerado e a elipse de confiança determinada no conjunto de treino é plotada nesse gráfico. Ao fim, é executada uma validação dos dados com relação aos pontos de dados dentro da elipse de confiança. A ramificação de FDA segue um fluxo semelhante.

Nas seções seguintes serão detalhados os modelos matemáticos apresentados no fluxo de dados mostrado na Figura 3.1.

3.4.1 PCA

Nesta seção é apresentado o algoritmo de PCA utilizado no conjunto de dados redimensionados via *StandardScaler*. Em um primeiro momento, um ajuste do modelo de PCA foi executado no conjunto de treino. Em seguida, tanto os dados de treino quanto de teste foram transformados utilizando o modelo previamente ajustado. PCA foi aplicada utilizando a função PCA do pacote *decomposition* da biblioteca *scikit-learn*.

3.4.1.1 Ajuste do modelo

Seja $\mathbf{X} \in \mathbb{R}^{n \times m}$ uma matriz de dados com m variáveis e n amostras, onde \mathbf{x}_i ($i = 1, 2, \dots, m$) são vetores com média igual a zero e variância igual a 1. PCA é ajustada em cima da matriz de covariância Σ desses dados, sendo calculada da seguinte forma:

$$\Sigma = \frac{1}{n-1} \mathbf{X}^T \mathbf{X} \quad (3.4)$$

Os autovetores da matriz Σ determinam as componentes principais e os

autovalores representam a magnitude correspondente. Assim, uma decomposição dessa matriz é feita:

$$\Sigma = \mathbf{P}\mathbf{\Lambda}\mathbf{P}^T \quad (3.5)$$

Onde \mathbf{P} representa a matriz de autovetores e $\mathbf{\Lambda}$ é uma matriz diagonal dos autovalores correspondentes. Os autovalores na PCA informam quanta variância pode ser explicada por seu autovetor associado, portanto, o maior autovalor indica que a maior variância nos dados foi observada na direção de seu autovetor. É possível calcular o percentual de variância explicada por cada componente da seguinte forma:

$$\frac{\lambda_j}{\sum_{i=1}^n \lambda_i} \quad (3.6)$$

Onde λ_j é a variância (autovalor) do componente principal j .

Do conjunto inicial de dados, m autovetores são obtidos. Esses autovetores são, então, ordenados de forma decrescente de acordo com os autovalores correspondentes. Esses autovetores ordenados são usados para transformar os dados.

3.4.1.2 Transformação dos dados

Após o ajuste do modelo nos dados normalizados de treino, uma transformação nos dados normalizados de treino e teste é executada.

$$\mathbf{t}_i = \mathbf{X}\mathbf{p}_i \quad (3.7)$$

Onde \mathbf{t}_i corresponde à componente principal i , \mathbf{X} corresponde a matriz $n \times m$ dos dados normalizados e \mathbf{p}_i corresponde ao autovetor i da matriz ordenada de autovetores determinada na seção anterior, $i = 1, 2$.

3.4.2 FDA

Assim como na Seção 3.4.1, esta seção é dividida em *ajuste do modelo* ao conjunto de treino e *transformação dos dados* do conjunto de treino e teste. Será

apresentado aqui o algoritmo de FDA utilizado nos conjuntos de dados redimensionados via *StandardScaler*. FDA foi aplicada utilizando a função *LinearDiscriminantAnalysis* do pacote *discriminant_analysis* da biblioteca *scikit-learn*.

3.4.2.1 Ajuste do modelo

Como levantado na Seção 2.2.2.6, a construção do modelo de FDA leva em consideração os dados de *input* e os dados categóricos de *output*. Assim, o modelo foi ajustado sob o conjunto de dados de *input* \mathbf{X} e um vetor \mathbf{y} de output categórico, representado pela variável CAT-LEVELS que apresenta 3 categorias, como mostrado na Tabela 3.3. Por ser mais conveniente trabalhar com dados numéricos, as categorias da variável CAT-LEVELS foram convertidas a inteiros, de forma que: 0 - Ótimo, 1 - Aceitável, 2 - Ruim.

O primeiro passo do modelo é computar os três vetores \mathbf{m}_i de dimensão m , que representam a média das variáveis em cada classe:

$$\mathbf{m}_i = \frac{1}{n_i} \sum_{x \in \text{classe } i}^n \mathbf{x}_k \quad (3.8)$$

Onde n_i representa a quantidade de amostras na classe i e \mathbf{x}_k representa o vetor de dimensão $1 \times m$ que compõe a matriz $\mathbf{X} \in \mathbb{R}^{n \times m}$.

Em seguida, a matriz de dispersão das classes é calculada:

$$\mathbf{S}_W = \sum_{i=0}^{c-1} \mathbf{s}_i \quad (3.9)$$

De forma que:

$$\mathbf{s}_i = \sum_{x \in \text{classe } i}^n (\mathbf{x}_k - \mathbf{m}_i) \cdot (\mathbf{x}_k - \mathbf{m}_i)^T \quad (3.10)$$

Depois, calcula-se a matriz de dispersão entre classes:

$$\mathbf{S}_B = \sum_{i=0}^{c-1} n_i \mathbf{m}_i \cdot \mathbf{m}_i^T \quad (3.11)$$

Então, são calculados os autovalores e autovetores da matriz $\mathbf{S}_W^{-1} \mathbf{S}_B$:

$$\mathbf{S}_W^{-1}\mathbf{S}_B = \mathbf{P}\mathbf{\Lambda}\mathbf{P}^T \quad (3.12)$$

Assim como na Equação 3.5, \mathbf{P} representa a matriz de autovetores e $\mathbf{\Lambda}$ representa a matriz diagonal de autovalores correspondentes. E, assim como no modelo de PCA, as componentes do modelo de FDA são determinadas a partir dos autovetores. Dessa forma, os autovetores são ordenados de forma decrescente de acordo com os autovalores correspondentes.

3.4.2.2 Transformação dos dados

Após o ajuste do modelo é realizada a transformação dos dados, para isso, é necessária apenas a matriz de dados (X). A transformação dos dados normalizados de treino e de teste é executada de forma semelhante à mostrada na Equação 3.7.

Ao fim, obteve-se para cada conjunto de dados dois vetores de componentes, \mathbf{t}_1 e \mathbf{t}_2 , contendo o máximo de informações possíveis a respeito do respectivo conjunto de dados.

3.4.3 Região Ótima

Para determinação de uma região que minimiza a quantidade de imperfeições presentes no produto gerado por aquele conjunto de variáveis de processo, os vetores de componentes principais obtidos pelo método de PCA para o conjunto de treino foram plotados em um gráfico 2D, o mesmo foi feito para os vetores obtidos por FDA para o conjunto de treino.

Para a determinação da região ótima, utilizou-se o método da elipse de confiança que será explicado mais à frente. Para a utilização desse método, é preciso fornecer o conjunto de pontos no qual a elipse deve ser plotada. Para a determinação desse conjunto de pontos, utilizou-se o método de clusterização de *KMeans*. Optou-se por *KMeans* por ser um método de clusterização não supervisionado de simples aplicação.

3.4.3.1 *KMeans*

O método de *KMeans* é um método de clusterização simples e foi executado seguindo os passos abaixo:

1. k centroides iniciais são escolhidos aleatoriamente;
2. cada ponto da distribuição é então atribuído ao centroide mais próximo, cada conjunto de pontos atribuídos a um mesmo centroide é chamado de *cluster*;
3. uma média dos valores de cada *cluster* é calculada;
4. os centroides são então atualizados como sendo a média dos seus respectivos *clusters*;
5. as etapas 2 a 4 são repetidas até que os centroides permaneçam os mesmos.

Deve ser destacado que o resultado final de *KMeans* depende da posição inicial dos centroides atribuídos de forma aleatória. Para sanar esse problema, *KMeans* é executado algumas vezes com diferentes posições iniciais, os resultados obtidos ao fim de cada rodada são comparados entre si e é retido aquele com menor inércia, que corresponde à soma das distâncias quadradas do ponto ao centro do seu próprio *cluster*.

Esse método foi aplicado utilizando a função *KMeans* do pacote *cluster* da biblioteca scikit-learn que tem como padrão 10 inicializações.

3.4.3.2 *Elipse de Confiança*

Após a determinação dos *clusters* via *KMeans*, o conjunto de pontos pertencentes ao *cluster* com menor quantidade média de imperfeições foi usado como base para a determinação da região de operação ótima via elipse de confiança. Essa elipse é construída a partir da matriz de covariância 2x2 de dados 2D. A estrutura dessa elipse pode ser observada na Figura 3.2.

Dados os vetores de componentes principais \mathbf{t}_1 e \mathbf{t}_2 dos dados de treino, a matriz de covariância é calculada de forma semelhante a Equação 3.4.

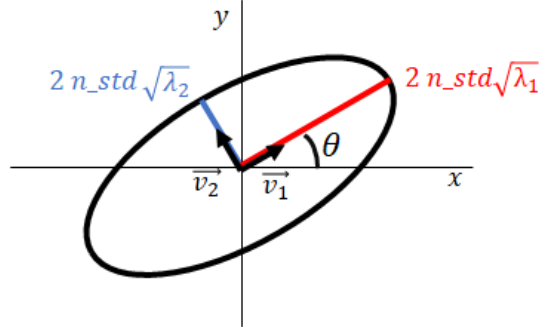


Figura 3.2: Elipse de confiança

$$\Sigma = \frac{1}{n-1} [\mathbf{t}_1, \mathbf{t}_2] [\mathbf{t}_1, \mathbf{t}_2]^t \quad (3.13)$$

Onde, n é a dimensão do vetor \mathbf{t}_i .

Os autovalores e autovetores dessa matriz são determinados pela Equação 3.5. Os autovetores são ordenados de acordo com os autovalores correspondentes. E as propriedades da elipse são determinadas da seguinte forma:

- Centro (x,y): é a média dos vetores \mathbf{t}_1 e \mathbf{t}_2 respectivamente;
- Largura: $2 n_std \sqrt{\lambda_1}$;
- Altura: $2 n_std \sqrt{\lambda_2}$;
- Ângulo: $\arctan\left(\frac{\mathbf{v}_{1x}}{\mathbf{v}_{1y}}\right)$

Onde λ_1 é o maior autovalor da matriz de covariância, λ_2 é o segundo maior autovalor da matriz de covariância, \mathbf{v}_1 é o autovetor com maior autovalor correspondente e n_std é o desvio padrão que determina o tamanho da elipse e é escolhido de modo que a elipse formada não inclua pontos ditos ruins no conjunto de treino.

3.5 Importância das Variáveis

Para a determinação da importância das variáveis nos modelos desenvolvidos, utilizou-se a biblioteca SHAP. SHAP é uma abordagem baseada na teoria dos

jogos para descrever o desempenho de um modelo de aprendizado de máquina. Para produzir um modelo interpretável, o SHAP usa um método de atribuição aditiva de variáveis, i.e., um modelo de saída é definido como uma adição linear de variáveis de entrada.

Dado um modelo de predição f que tem como *inputs* \mathbf{N} agentes, o valor Shapley que corresponde à contribuição de cada agente i na predição $f(\mathbf{N})$ será dado por:

$$\phi_i(f, \mathbf{N}) = \frac{1}{|\mathbf{N}|!} \sum_{\mathbf{S} \subseteq \mathbf{N} \setminus \{i\}} |\mathbf{S}|!(|\mathbf{N}| - |\mathbf{S}| - 1)! [f(\mathbf{S} \cup \{i\}) - f(\mathbf{S})] \quad (3.14)$$

\mathbf{S} corresponde a subconjuntos de amostras do conjunto inicial \mathbf{N} sem a amostra i ($\mathbf{N} \setminus \{i\}$). A subtração entre colchetes representa o cálculo marginal da amostra i , i.e., o que a amostra i afeta na predição $f(\mathbf{N})$. Esse cálculo é ponderado pela quantidade de possibilidades do cálculo marginal para a amostra i e, então, dividido pela quantidade de subconjuntos possíveis a partir do conjunto \mathbf{N} .

O valor Shapley satisfaz três condições desejáveis [17]:

- Simetria: Se i e j são tais que $f(\mathbf{S} \cup \{i\}) = f(\mathbf{S} \cup \{j\})$ para cada subconjunto \mathbf{S} , então $\phi_i(f, \mathbf{N}) = \phi_j(f, \mathbf{N})$;
- Jogador Nulo: quando a amostra não contribui para a predição, seu valor de contribuição é nulo;
- Linearidade: para duas funções de predição f_1 e f_2 na qual $(f_1 + f_2)(\mathbf{S}) = f_1(\mathbf{S}) + f_2(\mathbf{S})$ para todo subconjunto \mathbf{S} , então $\phi_i(f_1 + f_2, \mathbf{N}) = \phi_i(f_1, \mathbf{N}) + \phi_i(f_2, \mathbf{N})$;

A função *KernelExplainer* da biblioteca SHAP foi utilizada para explicar a saída de cada modelo desenvolvido e *shap_values* foi utilizado para se computar os valores Shapley.

4 Resultados e Discussões

Nesse Capítulo, serão discorridas análises e discussões acerca dos resultados obtidos com o desenvolvimento do trabalho.

4.1 Seleção de Variáveis

Como descrito na Seção 3.3.3, correlacionou-se as variáveis de processo com a variável de qualidade, *X3001*. O resultado pode ser observado na Figura 4.1. A média da correlação absoluta das variáveis é 0,168; as variáveis com coeficiente de correlação maior que 0,15 (valor próximo a média) foram previamente selecionadas.

Um mapa de calor da correlação entre as variáveis pré-selecionadas foi feito, Figura 4.2. Variáveis com valor absoluto de coeficiente de correlação maior do que 0,9 são variáveis altamente correlacionadas.

Inserir duas variáveis altamente correlacionadas no modelo não irá agregar novas informações. Com o objetivo de remover variáveis redundantes, analisou-se quais variáveis de processo estão duas a duas altamente correlacionadas, são elas:

- F0301R e F0301;
- Z1001SP e Z1001;
- T2401 e T2201;
- P2001O e S2001;

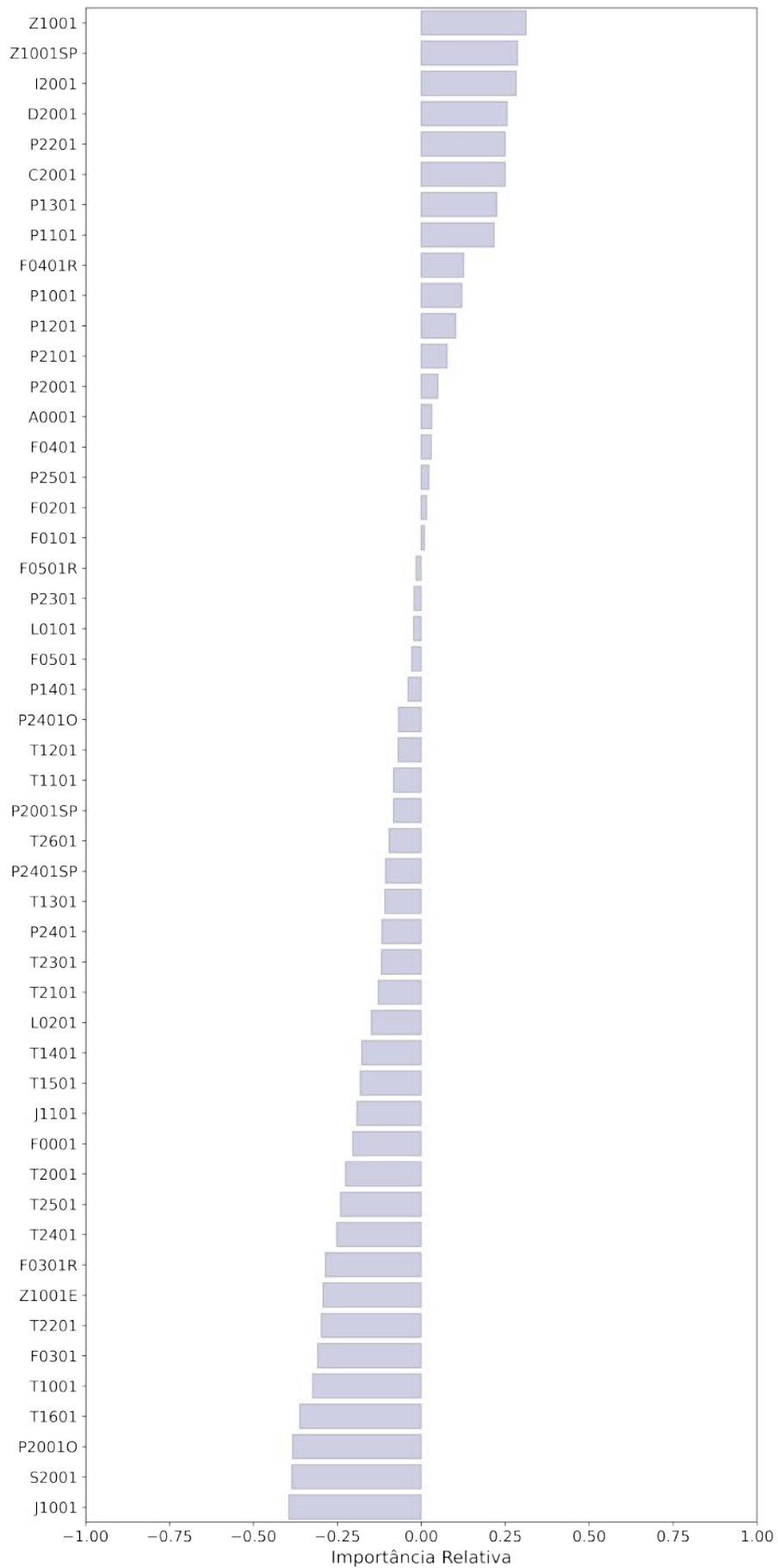


Figura 4.1: Correlação das variáveis de processos com a variável de quantidade de imperfeições

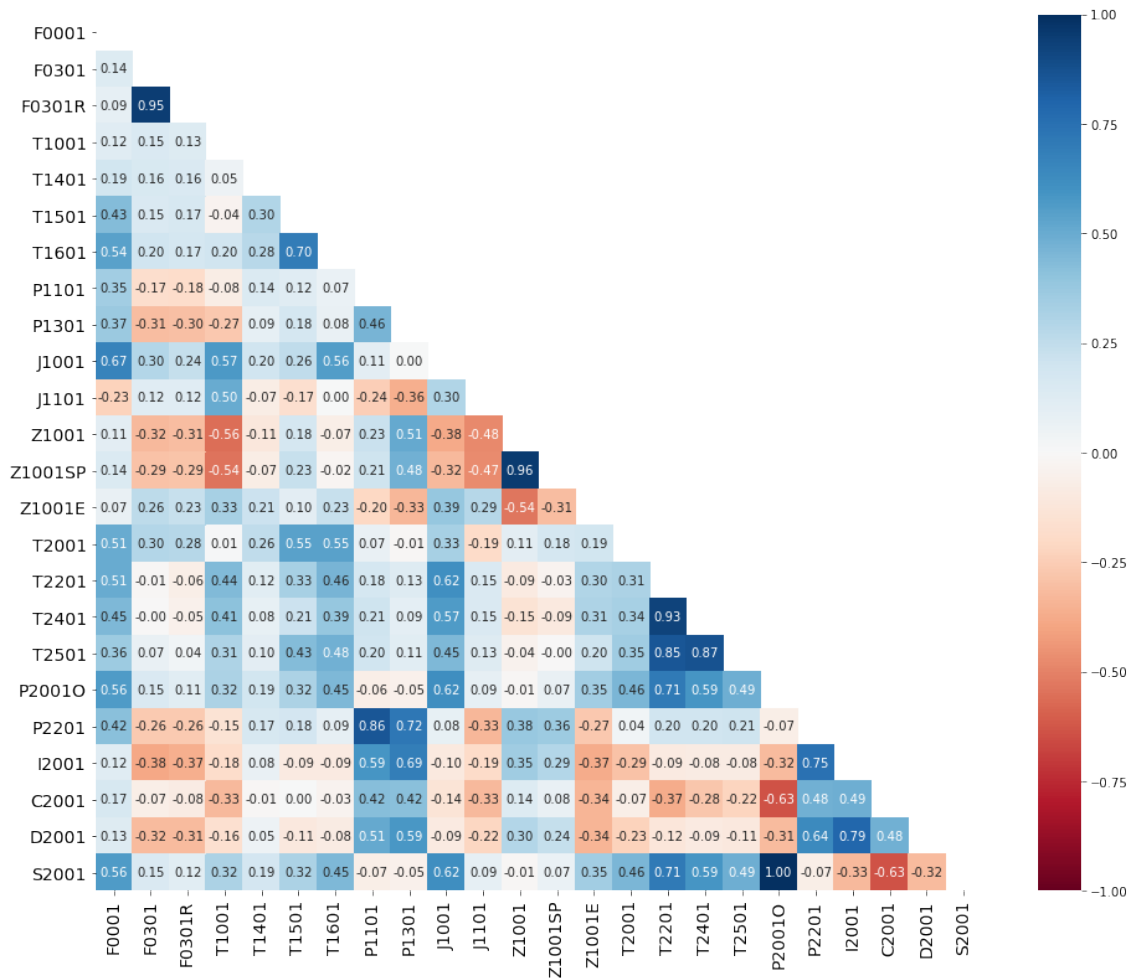


Figura 4.2: Mapa de calor da correlação das TAGs pré-selecionadas

Dos quatro pares de variáveis altamente correlacionadas, uma variável de cada par não foi inserida no modelo. Optou-se por remover a variável de cada par com menor valor de correlação em relação a variável de qualidade, $X3001$. Assim, ao final desse processo de seleção de variáveis, chegou-se a vinte variáveis de processo a serem usadas nos modelos desenvolvidos na sequência, são elas:

Tabela 4.1: TAGs selecionadas pelo método de Correlação

TAGs selecionadas
F0001
F0301
T1001
T1401
T1501
T1601

P1101
P1301
J1001
J1101
Z1001
Z1001E
T2001
T2201
T2501
P2201
I2001
C2001
D2001
S2001

Após a seleção das variáveis de processo, os dados brutos não normalizados foram divididos em um conjunto de treino e um conjunto de teste, conforme explicado na Seção 3.3.4. Para a construção dos modelos, o conjunto de treino seguiu o fluxo de dados mostrado na Figura 3.1, a validação foi feita no conjunto de teste, que seguiu o fluxo de dados de teste mostrado na mesma figura.

4.2 Determinação de regiões ótimas via PCA

4.2.1 Conjunto de Treino

Uma normalização via *Standard Scaler* foi ajustada e aplicada ao conjunto de treino. Seguindo o fluxo de dados, um modelo de PCA foi ajustado e aplicado a esses dados redimensionados.

A Figura 4.3 ajuda a entender como a variância é distribuída nas componentes principais. Esse é um importante gráfico que é utilizado para a determinação de quantas componentes serão selecionadas. Para se manter a explicabilidade dos modelos desenvolvidos, é desejável a seleção de duas componentes principais para ser possível a visualização 2D do problema. Observa-se que cerca de 53% de toda a variância do conjunto de dados está nas duas primeiras componentes, logo, selecionou-se as duas primeiras componentes principais para se prosseguir com o algoritmo.

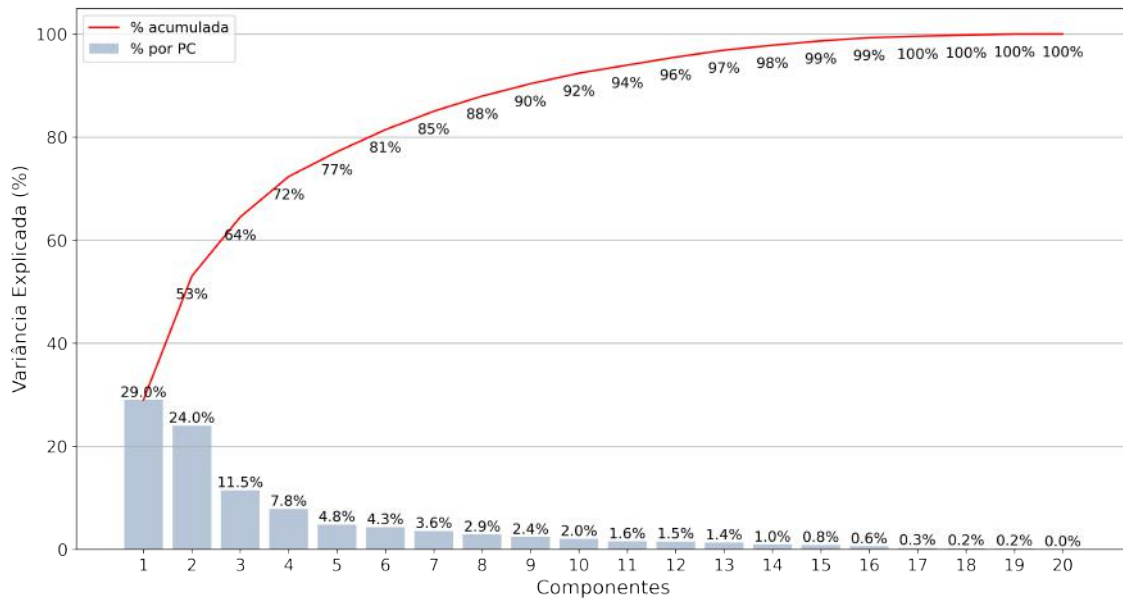


Figura 4.3: PCA - Variância Explicada e Acumulada das Componentes Principais

A Figura 4.4 traz a contribuição de cada variável para as componentes principais 1 e 2 do modelo de PCA ajustado. Observa-se que a variável mais importante para a componente 1 é a variável S2001 (rotação da zona 2) e as variáveis que mais contribuem para a componente 2 são a P2201 (pressão 2 da zona 2) e a F0001 (vazão de alimentação 0).

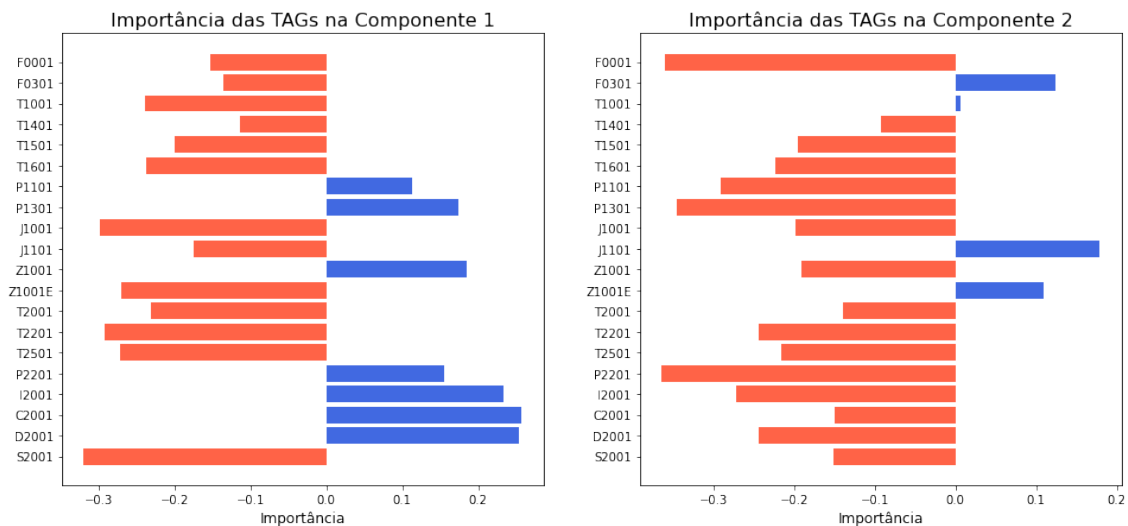


Figura 4.4: Importâncias das Variáveis sobre as Componentes da PCA

A Figura 4.5A corresponde ao gráfico da componente principal 1 e da componente principal 2 do conjunto de dados de treino, os pontos estão diferenciados de acordo com a variável CAT-LEVELS.

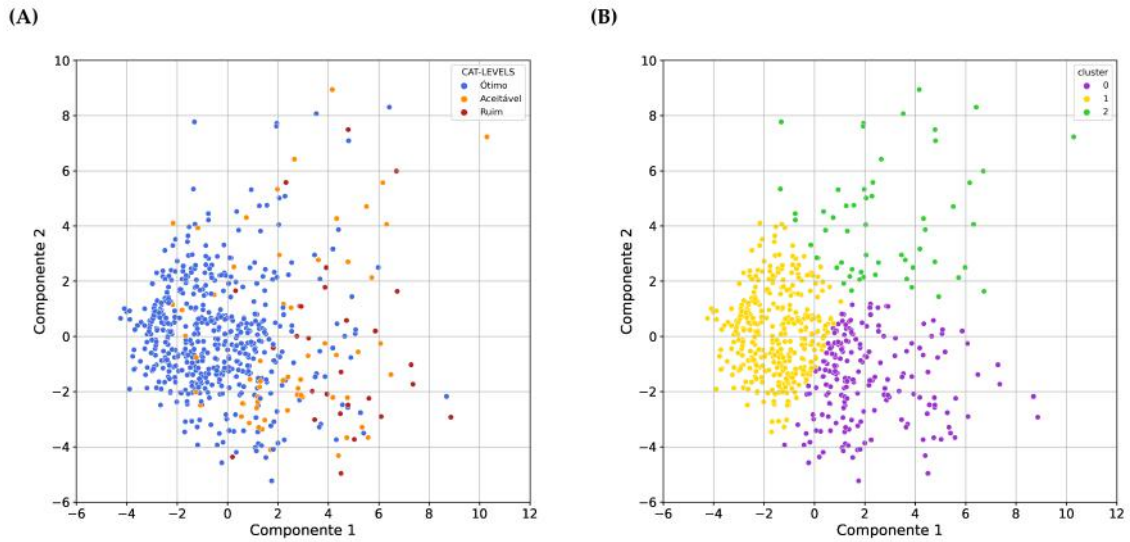


Figura 4.5: Gráfico 2D das componentes principais de PCA do conjunto de treino **(A)** discriminados pela variável categórica CAT-LEVELS e **(B)** discriminados pelos *clusters* determinados por *KMeans*.

Observa-se na Figura 4.5A que as três classes (Ótimo, Aceitável e Ruim) encontram-se misturadas. O objetivo aqui é determinar uma região no gráfico onde não existam pontos ruins e que contenha o máximo de pontos ótimos possível. Para a identificação dessa região, *KMeans* foi aplicado. Como explicado na Seção 3.4.3.1, *KMeans* não utiliza informações sobre a classe dos dados. Dessa forma, regiões bem definidas serão observadas. A Figura 4.5B mostra os *clusters* determinados por *KMeans*. Para a execução de *KMeans*, a seleção do número de *clusters* foi baseada na separação das regiões, observou-se que com três *clusters*, uma região com grande quantidade de pontos ótimos ficou bem definida, como é possível observar na imagem.

A quantidade média de imperfeições de cada *cluster* é exibida na Tabela 4.3. Como o *cluster* 1 foi aquele com menor quantidade de imperfeições, o conjunto de pontos pertencentes a esse *cluster* foi usado para se construir a elipse de confiança de acordo com a Seção 3.4.3.2.

A elipse final determinada a partir desse conjunto de treino via modelo de PCA é representada na Figura 4.6, o tamanho da elipse foi selecionado de modo que ela incluía o máximo de pontos ótimos e não continha pontos ruins. Observa-se que essa elipse inclui cerca de 58% de todos os pontos ótimos do conjunto de dados e não

Tabela 4.3: Quantidade média de imperfeições em cada *cluster* determinado por *KMeans* a partir do modelo via PCA

<i>cluster</i>	X3001 médio
0	93.6
1	51.1
2	107.9

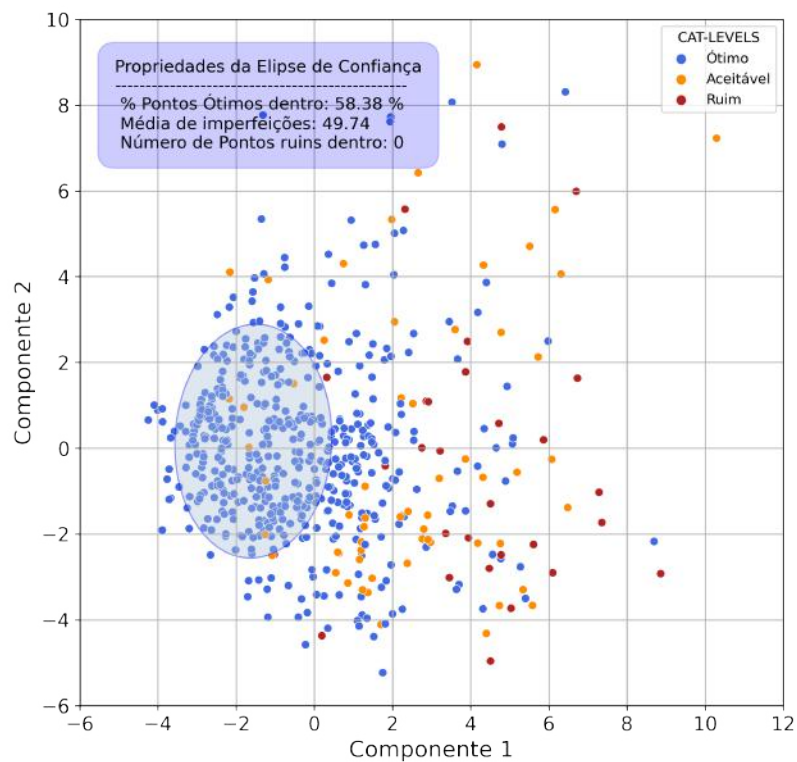


Figura 4.6: Gráfico 2D das componentes principais de PCA do conjunto de treino discriminados pela variável categórica CAT-LEVELS com elipse de confiança.

inclui nenhum ponto dito ruim. A quantidade média de imperfeições nessa região corresponde a média da quantidade de imperfeições dos pontos dentro dessa região, esse valor está abaixo da quantidade média de imperfeições dos pontos On-Spec (equivalente a 60,18).

4.2.2 Conjunto de Teste

O modelo ajustado ao conjunto de treino foi, então, aplicado ao conjunto de teste para se verificar sua capacidade de identificação de regiões operacionais que minimizam a quantidade de imperfeições em resinas poliméricas. O conjunto de dados de treino não normalizados seguiu o fluxo de dados mostrado na Figura 3.1.

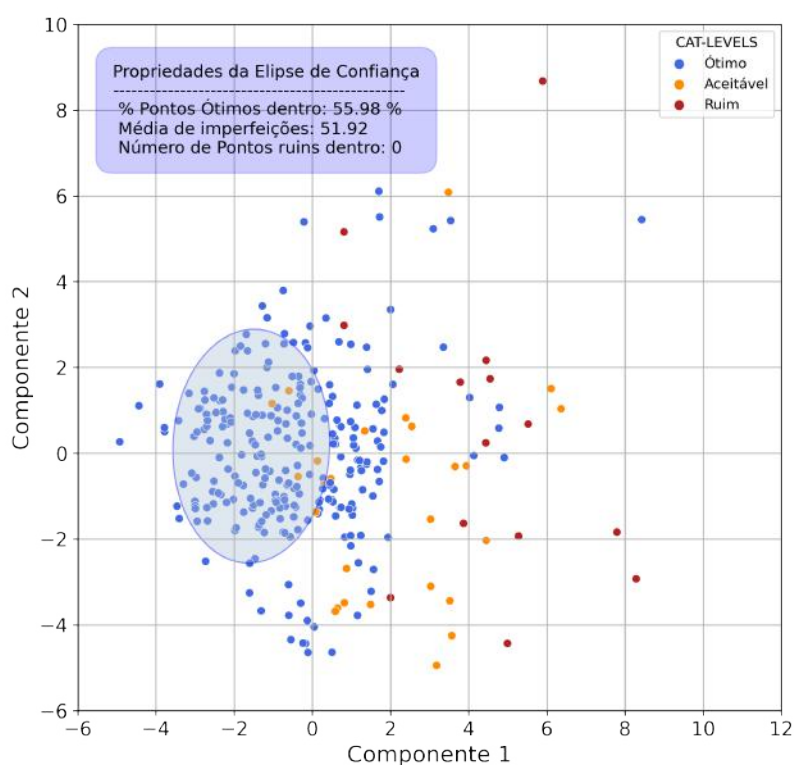


Figura 4.7: Gráfico 2D das componentes principais de PCA do conjunto de teste discriminados pela variável categórica CAT-LEVELS com elipse de confiança.

A Figura 4.7 apresenta o gráfico das componentes principais de PCA do conjunto de teste, juntamente com a elipse de confiança que foi determinada pelo conjunto de treino.

Observa-se que não há pontos ruins dentro dessa elipse e que a média da quantidade de imperfeições foi próxima a média da quantidade de imperfeições dentro da elipse do conjunto de teste. Assim, conclui-se que esse modelo desenvolvido via PCA é capaz de identificar uma região que minimiza a quantidade de imperfeições,

evitando a produção de resinas poliméricas *Off-Spec*.

4.3 Determinação de regiões ótimas via FDA

4.3.1 Conjunto de Treino

Como observado na Figura 3.1, a determinação de regiões ótimas via FDA seguiu uma modelagem semelhante à via PCA. A diferença entre as duas ramificações no fluxo de dados de treino é que na ramificação via FDA um modelo de FDA é ajustado ao invés de um modelo de PCA. A Figura 4.8 mostra a contribuição de cada variável para as componentes de FDA. Observa-se que a variável P2201 (pressão 2 da zona 2) foi a que mais contribuiu para a componente 1 e a S2001 (rotação da zona 2) a que mais contribuiu para a componente 2.

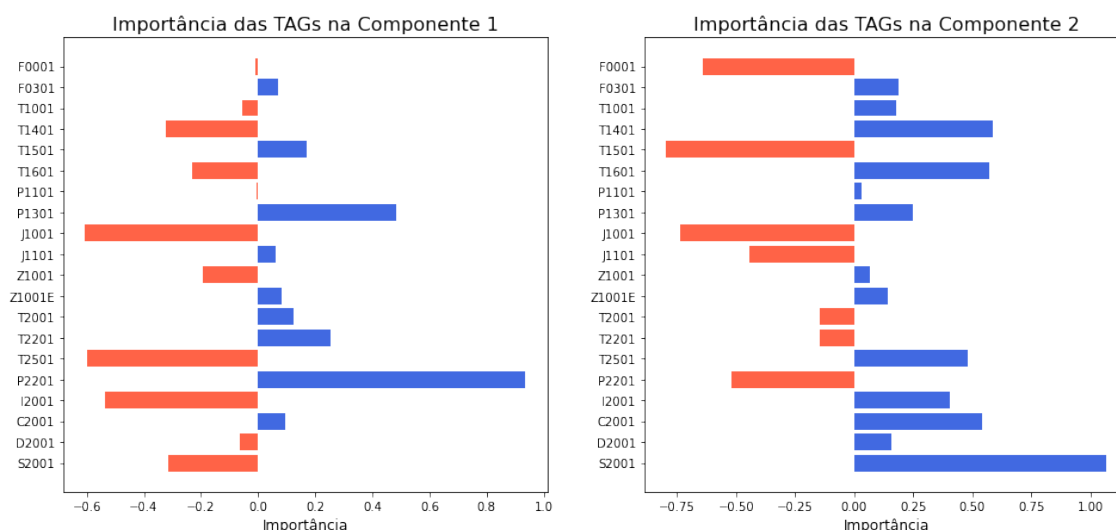


Figura 4.8: Importâncias das Variáveis sobre as Componentes da FDA

A Figura 4.9A corresponde ao gráfico da componente principal 1 e da componente principal 2 do conjunto de dados de treino do modelo de FDA, os pontos estão diferenciados pela variável CAT-LEVELS.

Semelhante ao resultado obtido pelo modelo via PCA na Figura 4.5A, as três classes (Ótimo, Aceitável e Ruim) não se encontram completamente separadas. Com o mesmo objetivo de determinar no gráfico da Figura 4.9A uma região em que não existam pontos ruins e que maximize a quantidade de pontos ótimos,

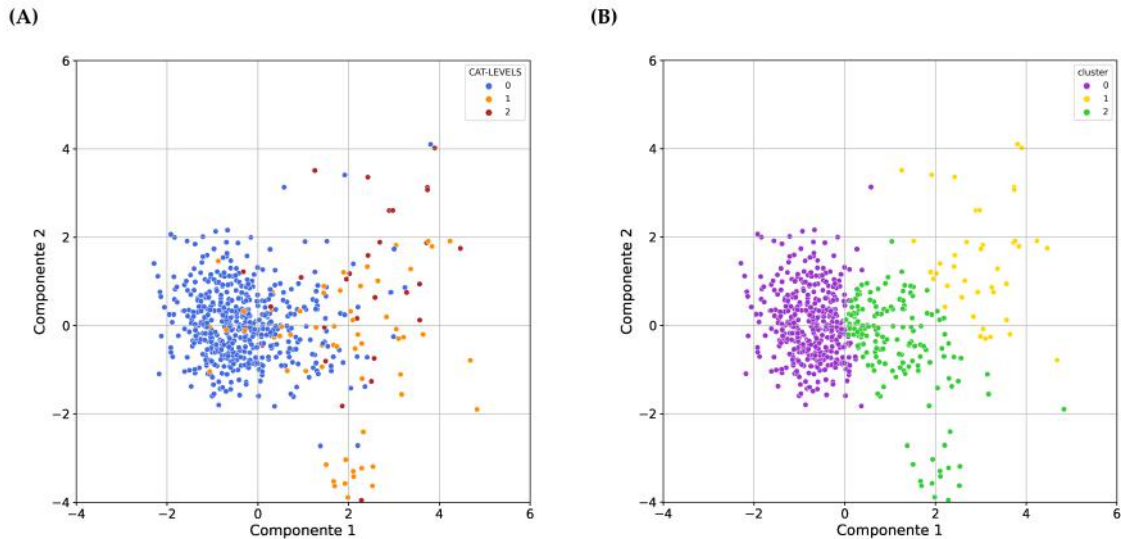


Figura 4.9: Gráfico 2D das componentes principais de FDA do conjunto de treino (A) discriminados pela variável categórica CAT-LEVELS e (B) discriminados pelos *clusters* determinados por *KMeans*.

utilizou-se *KMeans*. Os *clusters* determinados por *KMeans* podem ser observados na Figura 4.9B. Da mesma forma que feita pelo modelo via PCA, a seleção da quantidade de *clusters* baseou-se na separação da região visivelmente ótima, três *clusters* também foram selecionados dessa forma.

A quantidade média de imperfeições de cada *cluster* é exibida na tabela abaixo:

Tabela 4.4: Quantidade média de imperfeições em cada *cluster* determinado por *KMeans* a partir do modelo via FDA

<i>cluster</i>	X3001 médio
0	51.4
1	129.4
2	263.2

Da Tabela 4.4, observa-se que o *cluster* 0 é aquele com menor quantidade média de imperfeições, logo, o conjunto de pontos pertencentes a esse *cluster* foi usado para construir a elipse de confiança desse modelo. A elipse final determinada a partir desse conjunto de treino via modelo de FDA é representada na Figura 4.10.

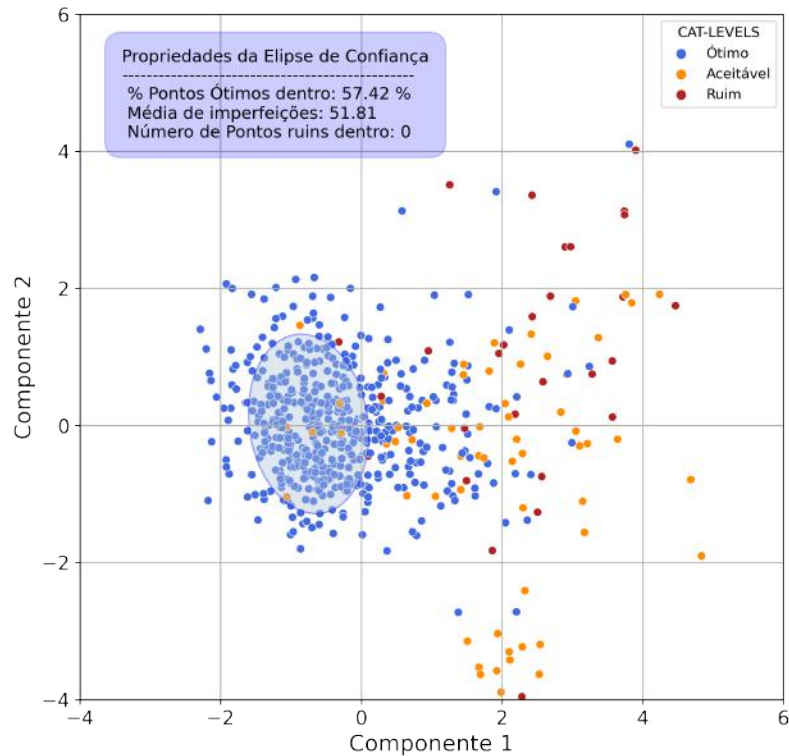


Figura 4.10: Gráfico 2D das componentes principais de FDA do conjunto de treino discriminados pela variável categórica CAT-LEVELS com elipse de confiança.

Essa elipse inclui cerca de 57.4% de todos os pontos ótimos do conjunto de dados de treino e não inclui nenhum ponto dito ruim. A quantidade média de imperfeições nessa região é de 51,8, valor abaixo da quantidade média de imperfeições dos pontos On-Spec (equivalente a 60,18), mostrando que a região delimitada determina pontos com baixa quantidade de imperfeições.

4.3.2 Conjunto de Teste

O modelo construído via FDA foi aplicado ao conjunto de dados de treino seguindo o fluxo de dados mostrados na Figura 3.1. A Figura 4.11 exibe o gráfico das componentes de FDA do conjunto de teste, juntamente com a elipse de confiança determinada pelo conjunto de treino.

Pelos dados observados dentro do quadro azul da Figura 4.11, pode-se inferir

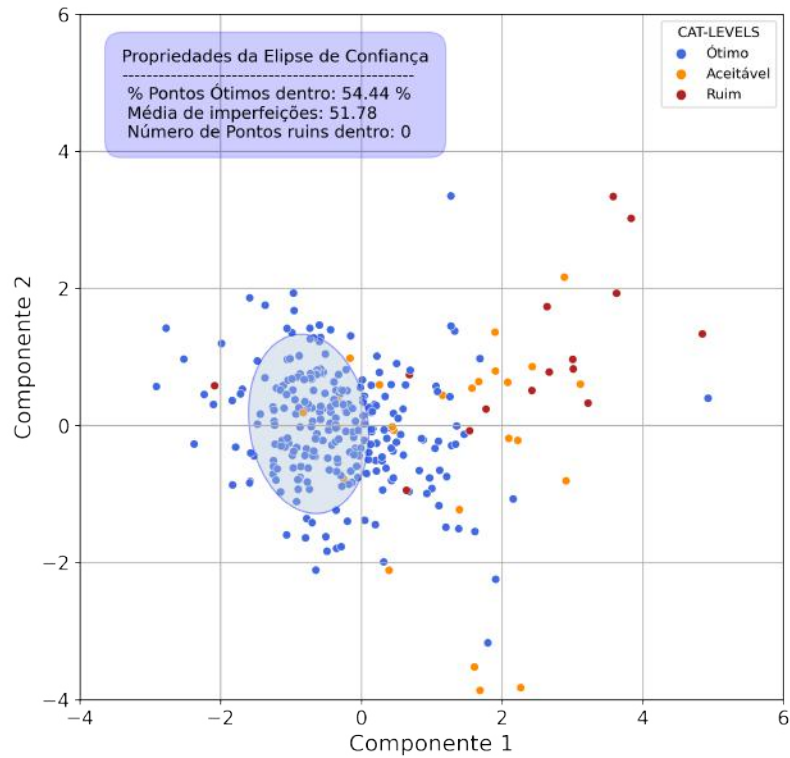


Figura 4.11: Gráfico 2D das componentes principais de FDA do conjunto de teste discriminados pela variável categórica CAT-LEVELS com elipse de confiança.

que o modelo desenvolvido via FDA também é capaz de identificar uma região onde evita-se pontos *Off-spec* e se minimiza a quantidade de imperfeições da resina polimérica.

4.4 Comparação dos modelos

Observa-se que tanto o modelo desenvolvido via PCA quanto o modelo desenvolvido via FDA foram capazes de atingir o resultado esperado: encontrar uma região densa que minimize a quantidade de imperfeições no produto polimérico produzido pelo equipamento industrial.

Realizando-se uma análise comparativa dos resultados obtidos, observa-se que mais de 70 % dos pontos presentes dentro da elipse definida pelo modelo via PCA estão presentes na elipse definida via FDA. De todos os dados disponíveis, o modelo via PCA incluiu 459 pontos *On-Spec*, já o modelo via FDA incluiu 449 pontos *On-Spec*. Observa-se que, em geral, o desempenho dos dois modelos foi semelhante, a média da quantidade de imperfeições dos pontos dentro da elipse determinada via PCA é próxima a média do modelo via FDA.

Com relação as diferenças estatísticas das variáveis do conjunto de pontos dentro e fora da elipse, é possível se observar histogramas das variáveis normalizadas dentro e fora da elipse determinada pelo modelo via PCA bem como o *boxplot* de cada uma das variáveis dentro e fora da elipse no ANEXO A. No ANEXO B é possível observar o mesmo para o modelo via FDA.

Nos histogramas, o eixo x representa o valor normalizado da *TAG*, já o eixo y representa a frequência de amostragem. Cada barra representa a quantidade de amostras que apresentam valor da *TAG* relacionada entre o mínimo e o máximo dessa barra no eixo x. Assim, é possível observar nesse tipo de gráfico a distribuição dos valores da *TAG* para o conjunto de pontos que se encontra dentro da elipse (barras azuis) e para o conjunto de pontos que se encontra fora da elipse (barras vermelhas). Dessa forma, é possível comparar a diferença entre essas duas distribuições.

O *boxplot* é uma outra forma de visualização da distribuição dos mesmos dados com um pouco mais de informações. A Figura 4.12 mostra as informações que são possíveis de se extrair de um *boxplot*. O eixo vertical representa o valor normalizado da *TAG* relacionada; o eixo horizontal representa o fator de interesse (DENTRO ou FORA da elipse); o quartil inferior (Q1) representa o valor o qual 25% das amostras possui um valor inferior; a mediana (Q2) é o valor onde metade

das amostras se encontram com valores acima e metade se encontram com valores abaixo; o quartil superior ($Q3$) é o valor o qual 25% das amostras tem valor superior a esse; o intervalo interquartilico (IQR) representa a diferença $Q3 - Q1$; o limite inferior corresponde a $Q1 - 1,5IQR$; o limite superior corresponde a $Q3 + 1,5IQR$; e *outliers* são amostras com valores menores que o limite inferior ou maiores que o limite superior.

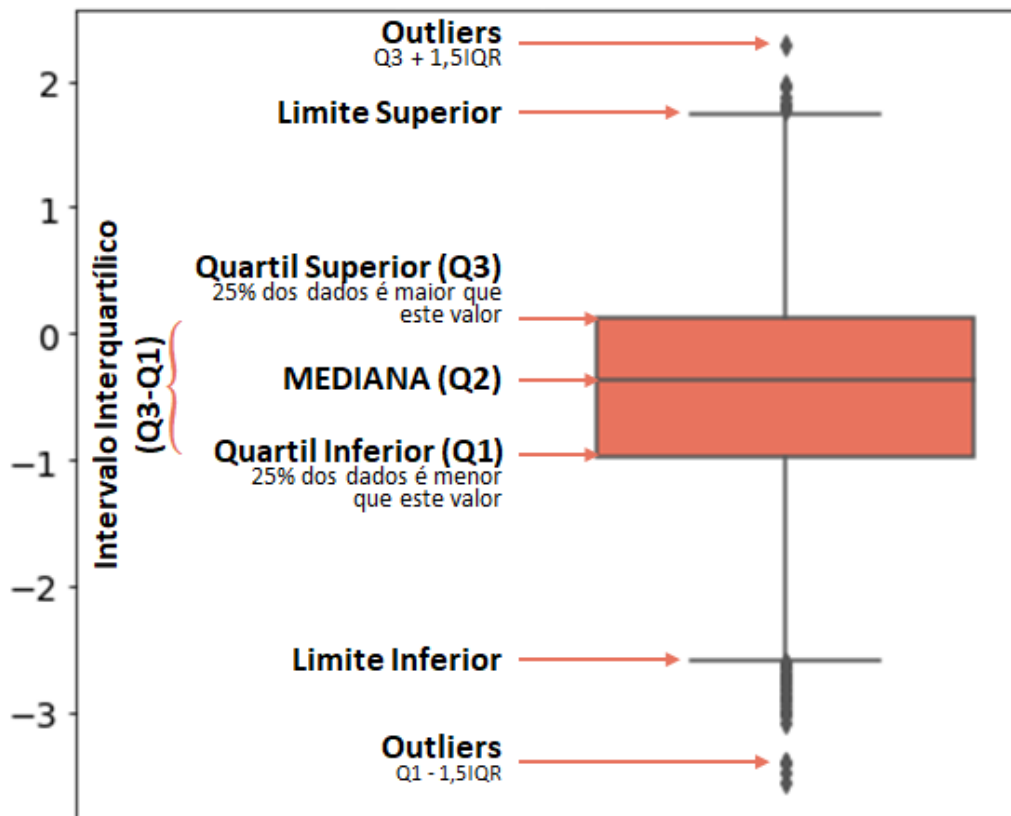


Figura 4.12: Elenco de informações contidas no *boxplot*

Através desses gráficos é possível se observar que a *TAG* S2001 (rotação da zona 2), nas páginas 62 e 69, é a *TAG* que apresenta distribuição mais distinta entre os dois grupos (DENTRO e FORA) para ambos os modelos. Ainda para ambos os modelos, a região ótima determinada apresenta valores de rotação mais altos que os valores para as amostras fora dessa região. A rotação relaciona-se a mistura dentro do equipamento, uma maior rotação assegura uma maior mistura e provavelmente uma menor quantidade de imperfeições no produto final.

Como discutido na Seção 2.4, uma das formas de aumentar a explicabilidade de modelos é via análise de valores *Shapley*. A Figura 4.13 mostra a média dos valores *Shapley* de cada variável para os dois modelos desenvolvidos. Observa-se que a importância das variáveis para cada modelo é bem similar a importância das *TAGs* na Componente 1, observadas na Figura 4.4A e na Figura 4.8A. Isso significa que, em ambos os modelos, a Componente 1 é a que possui maior influência na determinação da região ótima, fato que pode ser observado nos gráficos das elipses dos dois modelos.

Nota-se que a variável de maior importância para o modelo via PCA é a variável de rotação da zona 2 (S2001) que também é a variável que mais se notou diferença de distribuição nos gráficos em anexo, na página 62. Para o modelo de FDA, a variável de potência do equipamento (J1001) é a variável com maior importância e podemos observar nos gráficos de distribuição, na página 66, que os pontos dentro da elipse correspondem a valores mais altos de potência. Na Figura 4.2, é possível ver que a correlação da variável de potência e da variável de rotação da zona 2 é alta (0,62), logo, pode-se inferir disso que uma maior potência do equipamento corresponderá a uma maior rotação da zona 2.

Uma variável com alta importância atribuída para ambos os modelos é a variável de temperatura T2501. Essa temperatura relaciona-se a temperatura de elementos de transmissão mecânica, responsáveis por dar sustentação a um eixo que rotaciona. Essa temperatura está associada tanto a rotação desse eixo quanto a temperatura do produto. Nos gráficos de distribuição dessa *TAG*, nas páginas 61 e 68, é possível notar que quando as amostras estão dentro da elipse, essa temperatura está mais alta. Esse fato condiz com o esperado, já que a temperaturas mais altas a rotação está maior e/ou a temperatura do produto está maior e ambos esses fatores contribuem para a diminuição de imperfeições no produto final.

Foi notório que ambos os modelos desenvolvidos tiveram resultados semelhantes, logo, não é possível definir o melhor desses dois modelos com base nos

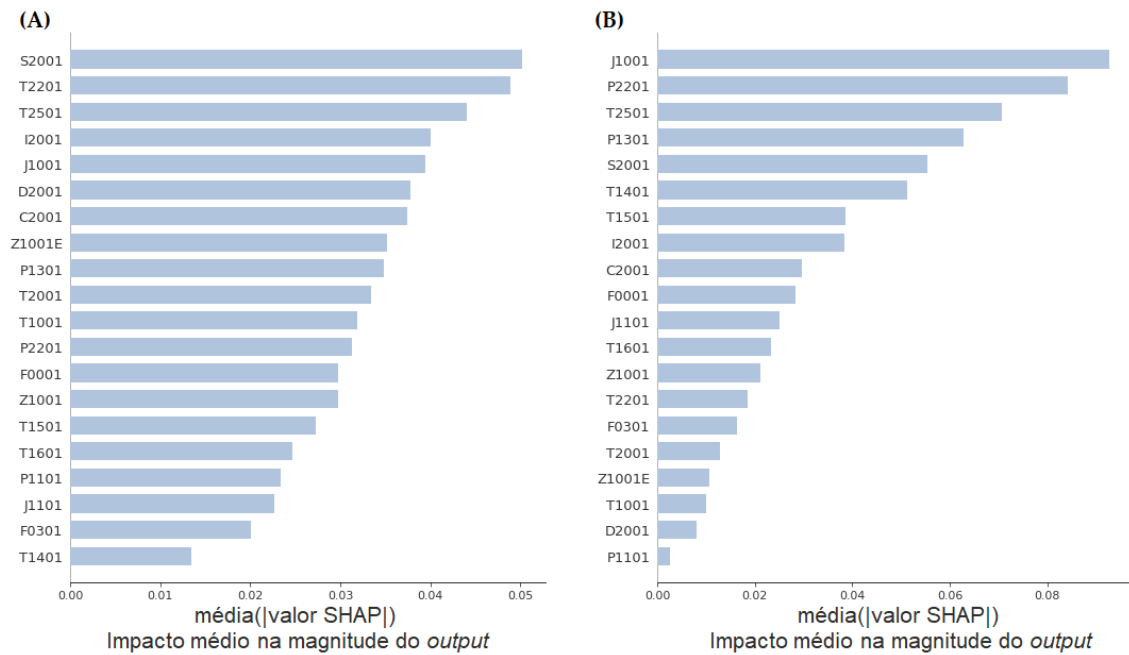


Figura 4.13: Contribuição das variáveis nos modelos desenvolvidos **(A)** via PCA e **(B)** via FDA

resultados para o caso estudado nesse projeto. Para este caso, por questões de simplicidade e maior quantidade de estudos a respeito, uma preferência maior pode ser dada ao modelo desenvolvido via PCA. Com esse estudo, foi possível obter informações importantes para a indústria. Foi possível determinar um *range* de operação que possibilita a minimização da quantidade de imperfeições no produto final, o que aumentará os lucros de produção. Esses *ranges* podem ser observados nos gráficos dos anexos, com os valores mínimos e máximos definidos pelo que cada variável assume na categoria DENTRO.

5 Conclusão

Com o avanço das tecnologias no meio industrial, cada vez mais dados são captados e armazenados. A análise de dados mostra-se como uma ferramenta muito interessante para a captura de informações relevantes que podem contribuir com o desenvolvimento dos processos industriais. Com o advento da inteligência artificial, o emprego de *Machine Learning* na indústria promove diversos benefícios como um alto controle da qualidade dos processos e produtos.

Neste trabalho, um estudo foi conduzido para a identificação de regiões que levam à minimização de imperfeições num produto polimérico gerado por um equipamento industrial. A análise da literatura sobre a detecção de anomalias em sistemas industriais foi o primeiro passo para a elaboração de um modelo adequado para o problema em questão.

Antes do desenvolvimento de modelos matemáticos para a aquisição do conhecimento desejado, os dados devem passar por um processamento para que os modelos tenham melhor desempenho. Em razão da grande quantidade de variáveis de monitoramento e controle existentes no equipamento industrial em análise, um processamento de dados executado nesse projeto foi a seleção de variáveis via método de correlação em relação à variável de qualidade (quantidade de imperfeições). Essa seleção permitiu reduzir a dimensionalidade, além de passar ao modelo as informações mais pertinentes.

Um fluxo de dados foi elaborado para a determinação das regiões desejadas. Os dados seguiram duas abordagens distintas, porém com o mesmo objetivo final, a primeira via modelo de Análise de Componente Principal e a segunda via modelo de Análise de Discriminante de Fisher. Ambas as abordagens tiveram como objetivo a redução da dimensionalidade do problema multivariado para um problema em duas dimensões.

Os dados iniciais foram separados em treino e teste para que fosse possível a validação dos modelos desenvolvidos. Os modelos foram construídos em cima da base de treino e, então, executados na base de teste.

É notável que ambos os modelos desenvolvidos foram capazes de atingir o objetivo final de identificação de uma região que conduz a menor quantidade de imperfeições presentes no produto final. Apesar das duas elipses construídas comportarem apenas cerca de 50 % do total de pontos ótimos (com quantidade baixa de imperfeições), nenhuma delas incluiu pontos *Off-Spec*, o que mostra que os modelos desenvolvidos foram capazes de definir uma região que é susceptível a formação de pequenas quantidades de imperfeições.

Valores Shapley foram usados para se avaliar a importância das variáveis de cada modelo. Notou-se que ambos os modelos tiveram importância de variáveis bem similares, o que é algo esperado já que os resultados de ambos os modelos são bem próximos. A variável com maior importância apontada é a variável de rotação que é um resultado que pode ser explicado, pois a rotação pode ser diretamente relacionada a mistura dentro do equipamento.

Em questão de resultados, não é possível apontar um modelo como sendo o melhor. Entretanto, pelo fato da PCA apresentar uma quantidade maior de informações na literatura, pode-se escolher esse modelo para um maior aprofundamento na análise de dados.

É de grande valia destacar também que o projeto desenvolvido pôde definir um *range* operacional das variáveis de processo que possibilitam a minimização da quantidade de imperfeições formadas no produto polimérico em questão. Conhecimentos de processo, em geral, concentram-se em um conjunto pequeno de variáveis. Porém, para garantir a qualidade do produto, é necessário manter um grande conjunto de variáveis dentro de um *range* adequado, e essa informação foi obtida através da análise de dados e dos modelos aqui desenvolvidos. Assim, o projeto desenvolvido através do uso de uma grande quantidade de dados obteve informações importantes para os tomadores de decisão, permitindo um aumento da qualidade do produto, diminuição da produção de produto *Off-spec* e, por consequência, um aumento nos lucros.

Como perspectivas desse projeto, destaca-se o uso de outros métodos de redução de dimensionalidade, bem como de clusterização. PLS é apontado na literatura científica como um bom método para detecção de qualidade de produtos em

processos industriais e a utilização desse método pode trazer resultados interessantes a serem avaliados. Outra possibilidade é aumentar a dimensionalidade do espaço de busca. Neste trabalho, optou-se por resultados 2D para maior explicabilidade dos modelos desenvolvidos, mas uma investigação de espaços de mais dimensões pode trazer resultados mais precisos a custo da perda de interpretabilidade.

REFERÊNCIAS

- [1] BOTRE, C., MANSOURI, M., NOUNOU, M., NOUNOU, H., E KARIM, M. N. Kernel pls-based glrt method for fault detection of chemical processes. *Journal of Loss Prevention in the Process Industries* 43 (2016), 212–224.
- [2] BRONSHTEIN, A. Towards Data Science train/test split and cross validation in python. <https://towardsdatascience.com/train-test-split-and-cross-validation-in-python-80b61beca4b6>, 2017. Acessado em 09/06/2021.
- [3] BROWNLEE, J. Machine Learning Mastery data leakage in machine learning. <https://machinelearningmastery.com/data-leakage-machine-learning/>, 2016. Acessado em 10/06/2021.
- [4] DAI, Y., CHENG, F., WU, H., WU, D., E ZHAO, J. Chapter five - data driven methods. In *Advanced Methods of Risk Assessment and Management*, F. I. Khan e P. R. Amyotte, Eds., vol. 4 of *Methods in Chemical Process Safety*. Elsevier, 2020, pp. 167–203.
- [5] DING, S. X. Chapter seven - residual generation with enhanced robustness against unknown inputs. In *Model-based Fault Diagnosis Techniques: Design Schemes, Algorithms, and Tools*, S. X. Ding, Ed. Springer, 2008, pp. 161–246.
- [6] DING, S. X. *Data-driven Design of Fault Diagnosis and Fault-tolerant Control Systems*. Advances in Industrial Control. Springer, 2014.
- [7] GHOSH, K., RAMTEKE, M., E SRINIVASAN, R. Optimal variable selection for effective statistical process monitoring. *Computers & Chemical Engineering* 60 (2014), 260–276.
- [8] ISERMANN, R. Model-based fault-detection and diagnosis – status and applications. *Annual Reviews in Control* 29, 1 (2005), 71–85.
- [9] ISERMANN, R. *Fault-Diagnosis Systems: An Introduction from Fault Detection to Fault Tolerance*, 1 ed. Springer-Verlag Berlin Heidelberg, 2006.

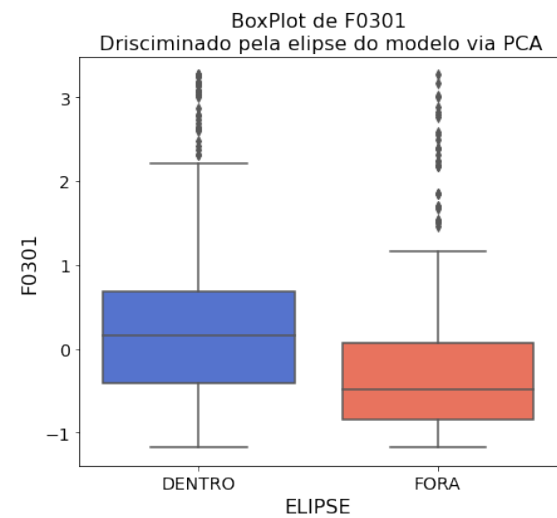
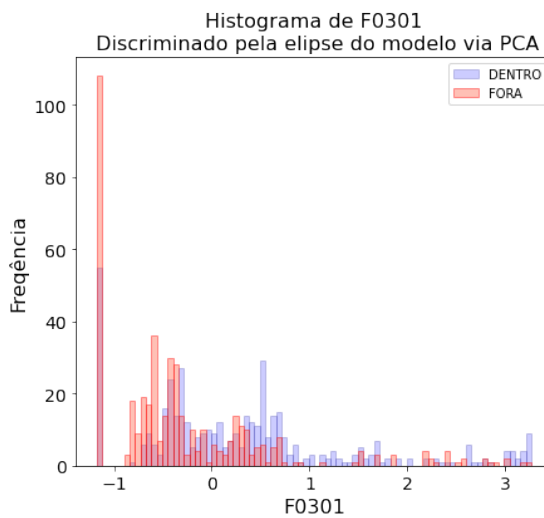
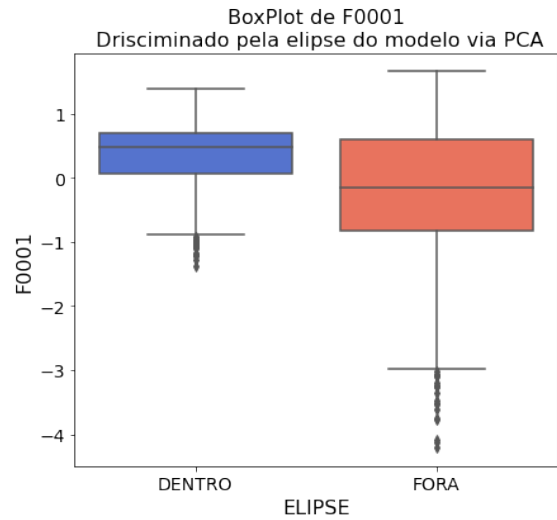
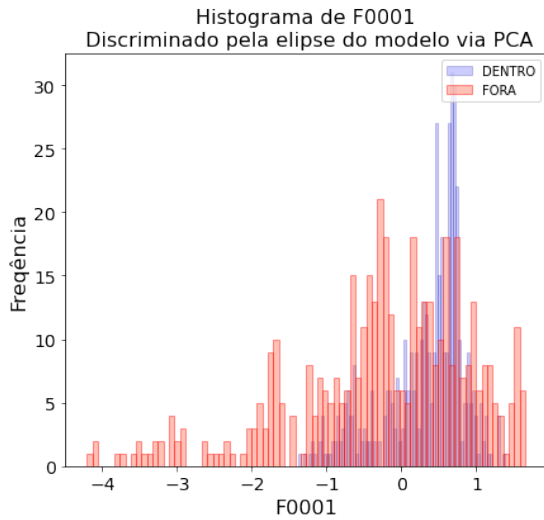
- [10] JAFARZADEGAN, M., SAFI-ESFAHANI, F., E BEHESHTI, Z. Combining hierarchical clustering approaches using the pca method. *Expert Systems with Applications 137* (2019), 1–10.
- [11] JAMAL, A., HANDAYANI, A., SEPTIANDRI, A., RIPMIATIN, E., E EFFENDI, Y. Dimensionality reduction using pca and k-means clustering for breast cancer prediction. *Lontar Komputer : Jurnal Ilmiah Teknologi Informasi* (12 2018), 192.
- [12] JULIER, S. J., E UHLMANN, J. K. New extension of the Kalman filter to non-linear systems. In *Signal Processing, Sensor Fusion, and Target Recognition VI* (1997), I. Kadar, Ed., vol. 3068, International Society for Optics and Photonics, SPIE, pp. 182 – 193.
- [13] LARRAÑAGA, P., ATIENZA, D., DIAZ-ROZO, J., OGBECHIE, A., PUERTO-SANTANA, C., E BIELZA, C. *Industrial Applications of Machine Learning*. CRC Press, 2018.
- [14] LEE, J.-M., QIN, S. J., E LEE, I.-B. Fault detection and diagnosis based on modified independent component analysis. *AIChE Journal 52*, 10 (2006), 3501–3514.
- [15] LI, Q., LI, R., JI, K., E DAI, W. Kalman filter and its application. In *2015 8th International Conference on Intelligent Networks and Intelligent Systems (ICINIS)* (2015), pp. 74–77.
- [16] LO, N. G., FLAUS, J.-M., E ADROT, O. Review of machine learning approaches in fault diagnosis applied to iot systems. In *2019 International Conference on Control, Automation and Diagnosis (ICCAD)* (2019), pp. 1–6.
- [17] MANGALATHU, S., HWANG, S.-H., E JEON, J.-S. Failure mode and effects analysis of rc members based on machine-learning-based shapley additive explanations (shap) approach. *Engineering Structures 219* (2020).
- [18] MILJKOVIĆ, D. Fault detection methods: A literature survey. In *2011 Proceedings of the 34th International Convention MIPRO* (2011), pp. 750–755.

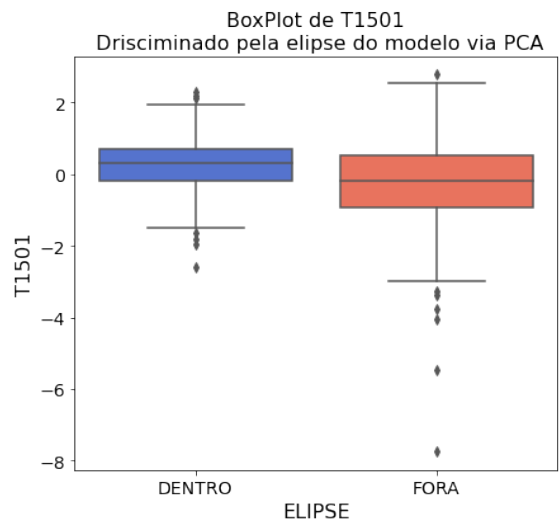
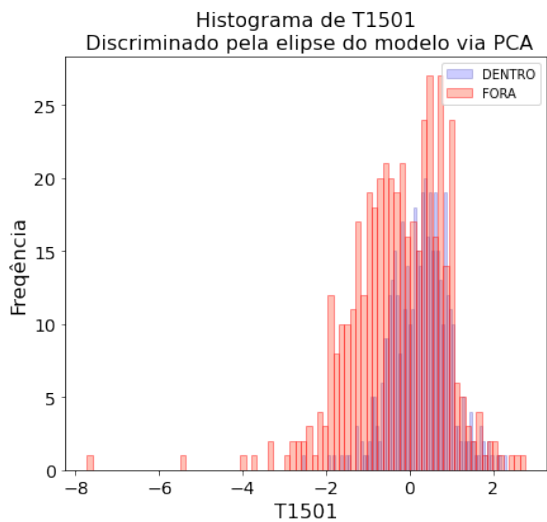
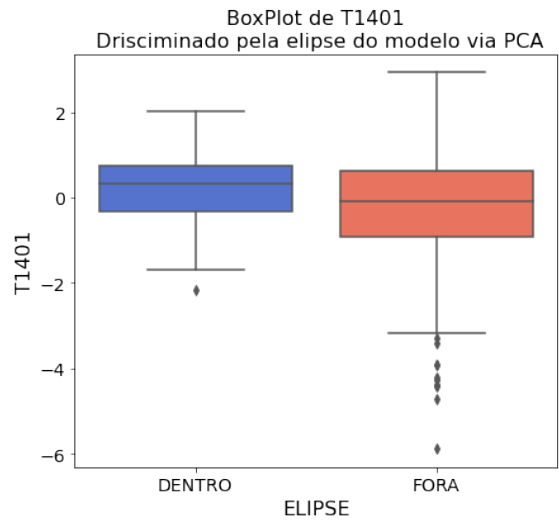
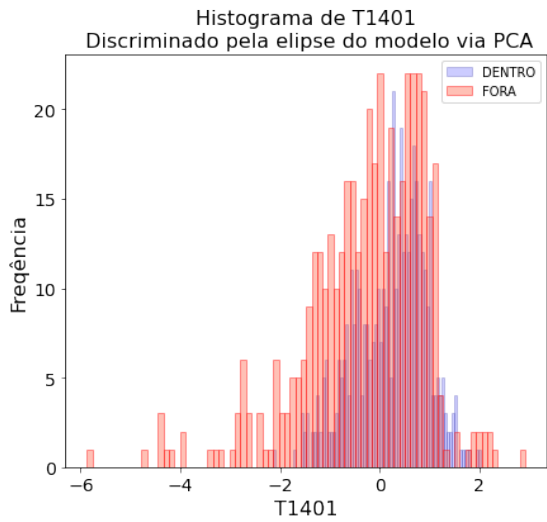
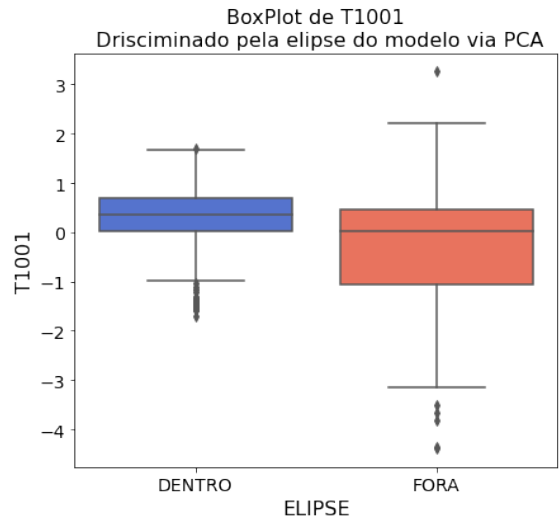
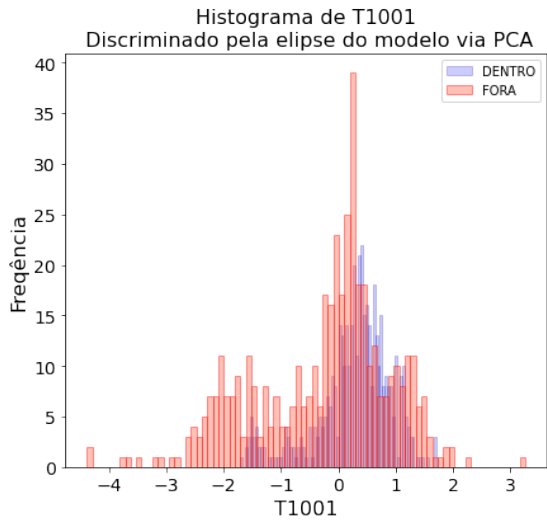
- [19] MINITAB, L. A comparison of the pearson and spearman correlation methods. <https://support.minitab.com/en-us/minitab-express/1/help-and-how-to/modeling-statistics/regression/supporting-topics/basics/a-comparison-of-the-pearson-and-spearman-correlation-methods/#:~:text=The%20Pearson%20correlation%20evaluates%20the%20linear%20relationship%20between%20two%20continuous%20variables.&text=The%20Spearman%20correlation%20coefficient%20is,evaluate%20relationships%20involving%20ordinal%20variables.,> 2019.
- [20] MUJICA, L., RODELLAR, J., GUEMES, A., E LÓPEZ-DIEZ, J. Pca based measures: Q-statistic and t2-statistic for assessing damages in structures. *Proceedings of the 4th European Workshop on Structural Health Monitoring* (01 2008), 1088–1095.
- [21] NOR, N. M., HASSAN, C. R. C., E HUSSAIN, M. A. A review of data-driven fault detection and diagnosis methods: applications in chemical process systems. *Reviews in Chemical Engineering* 36, 4 (2020), 513–553.
- [22] RATHI, P. Towards Data Science a novel approach to feature importance — shapley additive explanations. <https://towardsdatascience.com/a-novel-approach-to-feature-importance-shapley-additive-explanations-d18af30fc21b>, 2020. Acessado em 19/07/2021.
- [23] RODRÍGUEZ-PÉREZ, R., E BAJORATH, J. Interpretation of machine learning models using shapley values: application to compound potency and multi-target activity predictions. *Journal of Computer-Aided Molecular Design* 34 (2020), 1013–1026.
- [24] RUSSELL, E. L., CHIANG, L. H., E BRAATZ, R. D. *Data-driven Methods for Fault Detection and Diagnosis in Chemical Processes*. Spring, 2000.
- [25] SARTINI, B. A., GARBUGIO, G., BORTOLOSSI, H. J., SANTOS, P. A., E BARRETO, L. S. *Uma Introdução a Teoria dos Jogos*. II Bienal da SBM - Universidade Federal da Bahia, 2004.
- [26] SCHWAB, K. *The Fourth Industrial Revolution*. Crown Publishing Group, USA, 2017.

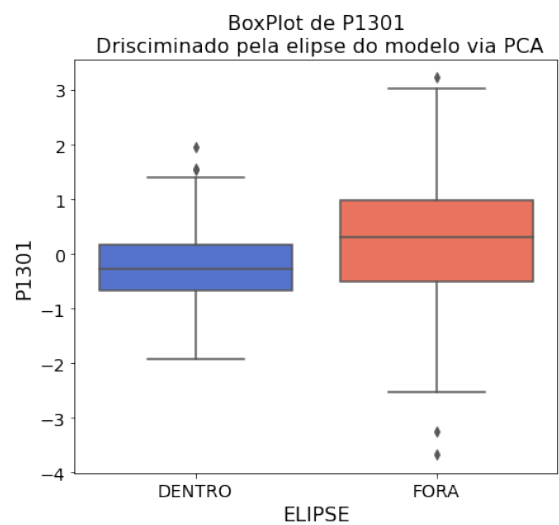
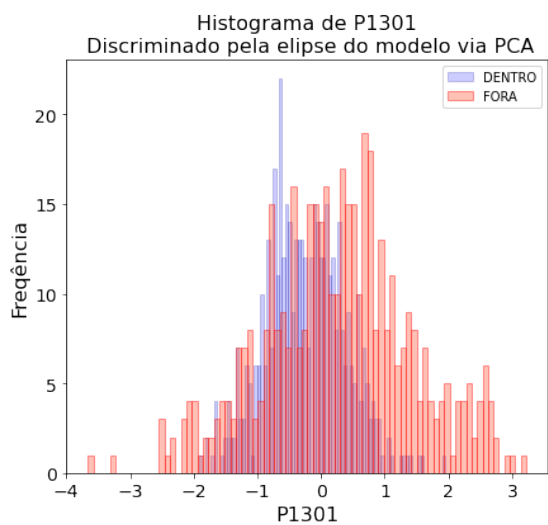
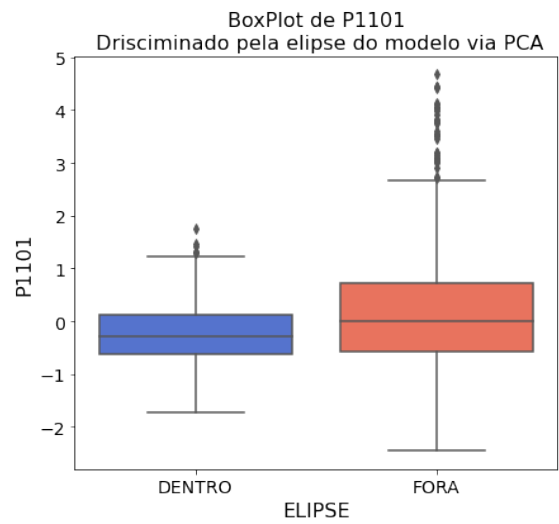
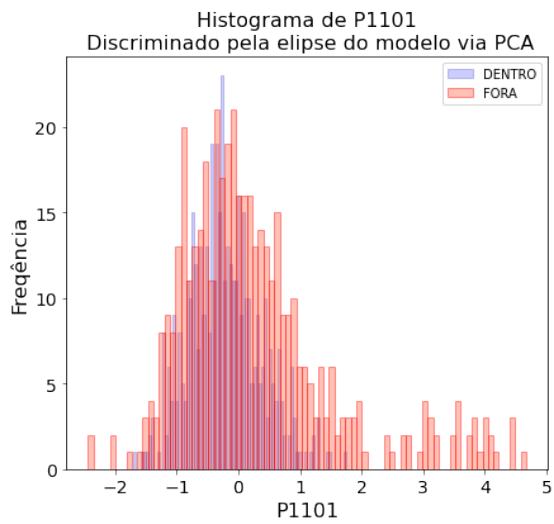
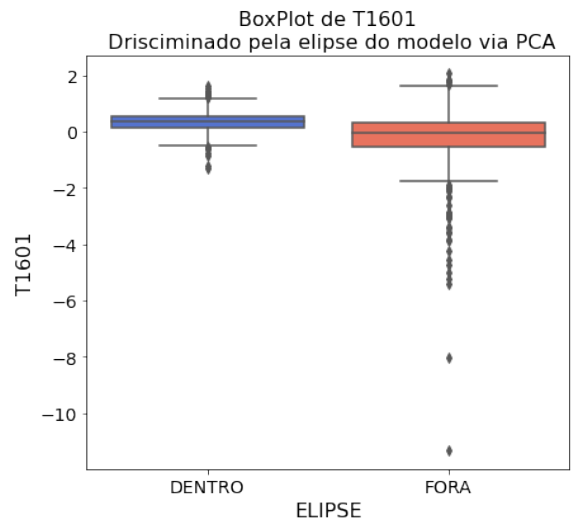
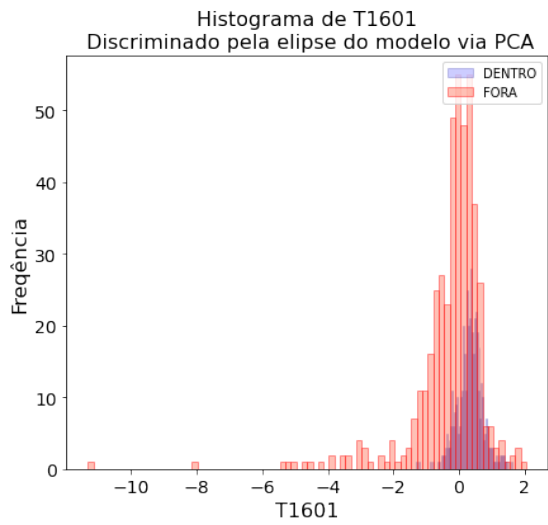
- [27] SHEARER, C. The crisp-dm model: The new blueprint for data mining. *Journal of Data Warehousing* 5, 4 (2000), 13–22.
- [28] TAN, F., FU, X., ZHANG, Y., E BOURGEOIS, A. A genetic algorithm-based method for feature subset selection. *Soft Comput.* 12 (01 2008), 111–120.
- [29] TAO, E., SHEN, W., LIU, T., E CHEN, X. Fault diagnosis based on pca for sensors of laboratorial wastewater treatment process. *Chemometrics and Intelligent Laboratory Systems* 128 (10 2013), 49–55.
- [30] VENKATASUBRAMANIAN, V., RENGASWAMY, R., E KAVURI, S. N. A review of process fault detection and diagnosis: Part ii: Qualitative models and search strategies. *Computers & Chemical Engineering* 27, 3 (2003), 313–326.
- [31] VENKATASUBRAMANIAN, V., RENGASWAMY, R., KAVURI, S. N., E YIN, K. A review of process fault detection and diagnosis: Part iii: Process history based methods. *Computers & Chemical Engineering* 27, 3 (2003), 327–346.
- [32] VENKATASUBRAMANIAN, V., RENGASWAMY, R., YIN, K., E KAVURI, S. N. A review of process fault detection and diagnosis: Part i: Quantitative model-based methods. *Computers & Chemical Engineering* 27, 3 (2003), 293–311.
- [33] YIN, S., DING, S., HAGHANI, A., HAO, H., E ZHANG, P. A comparison study of basic data-driven fault diagnosis and process monitoring methods on the benchmark tennessee eastman process. *Journal of Process Control* 22 (10 2012), 1567–1581.

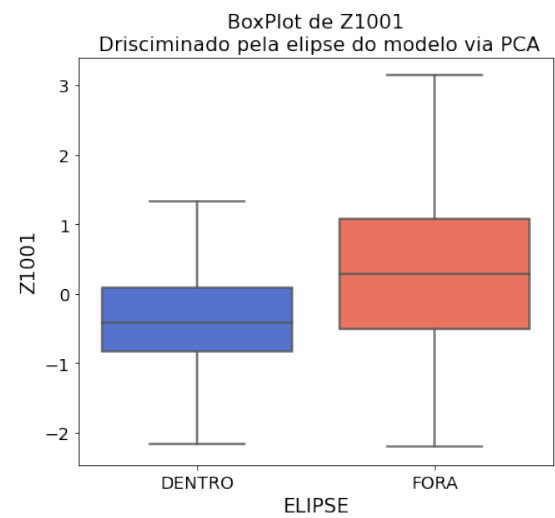
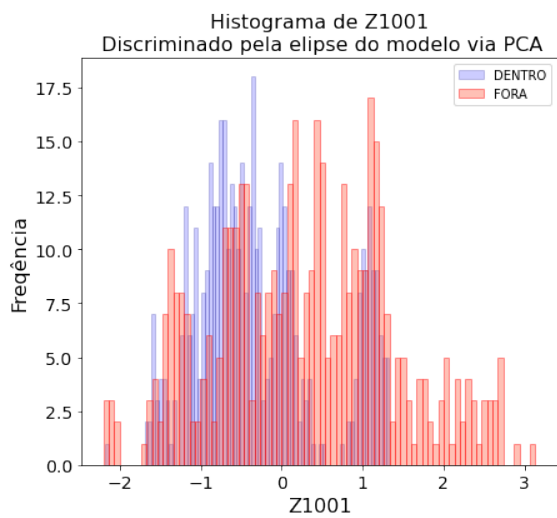
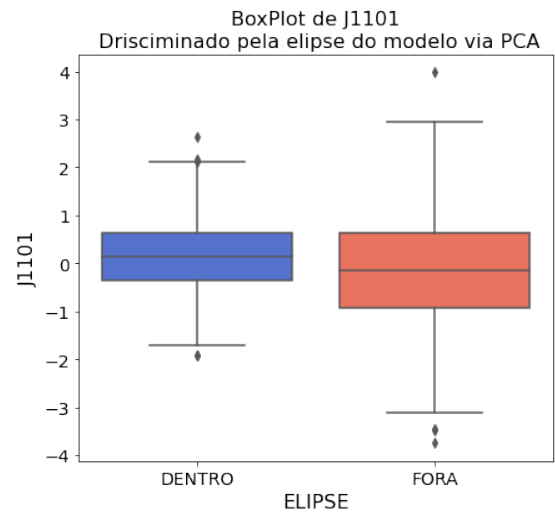
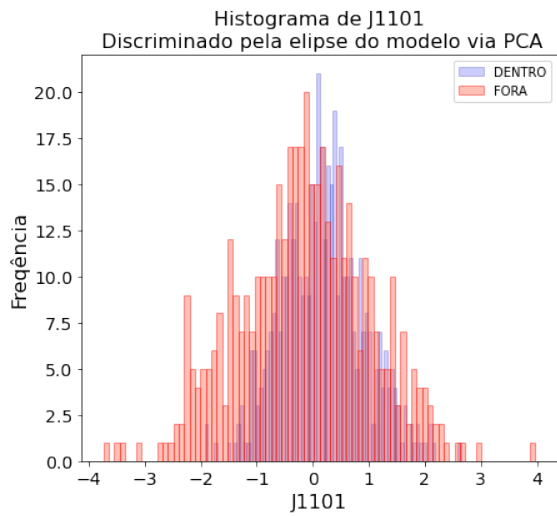
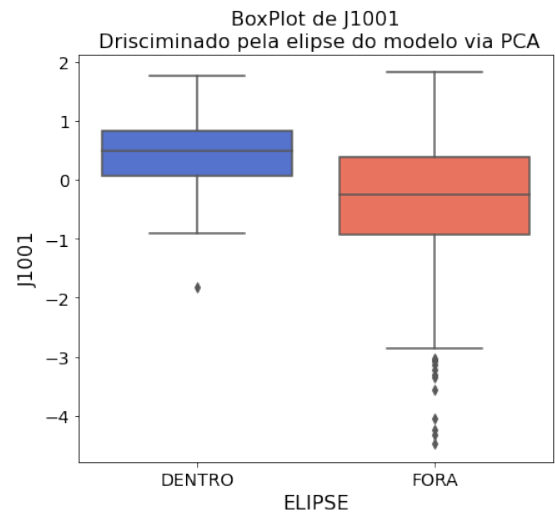
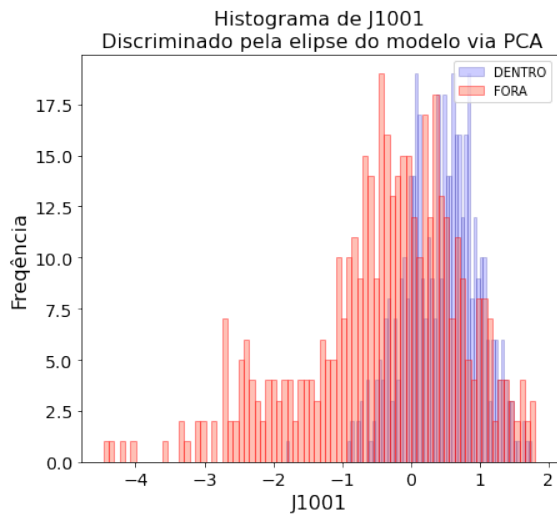
ANEXOS

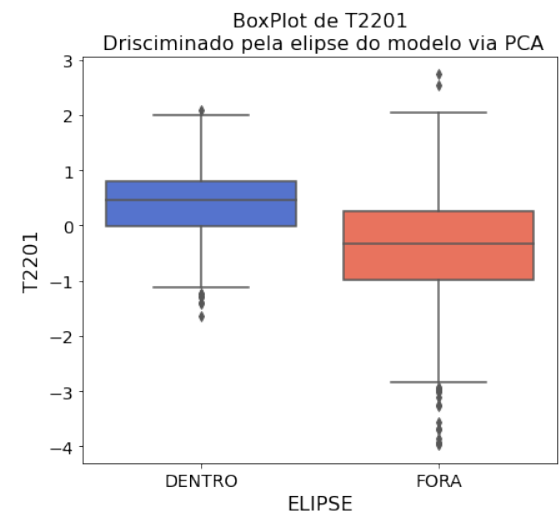
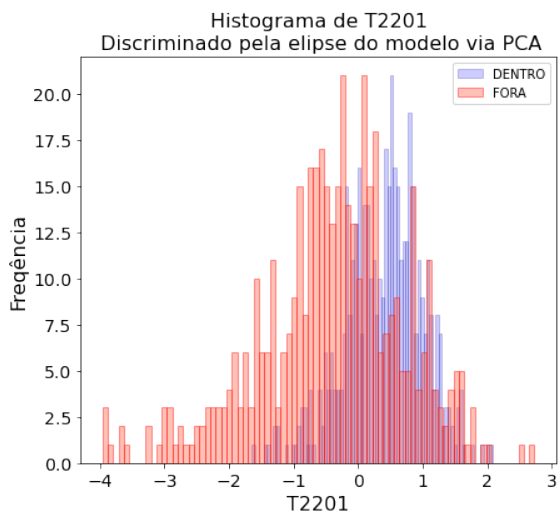
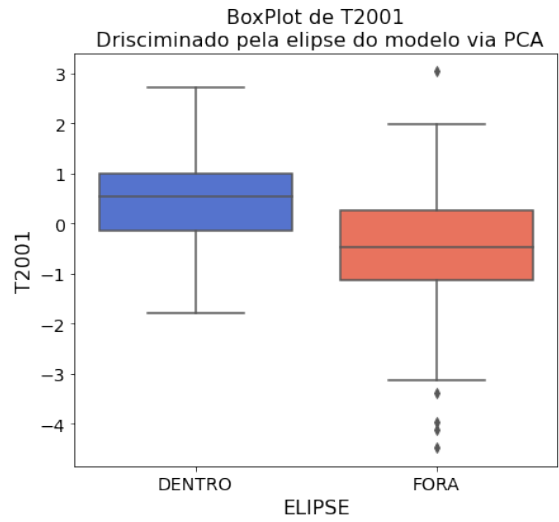
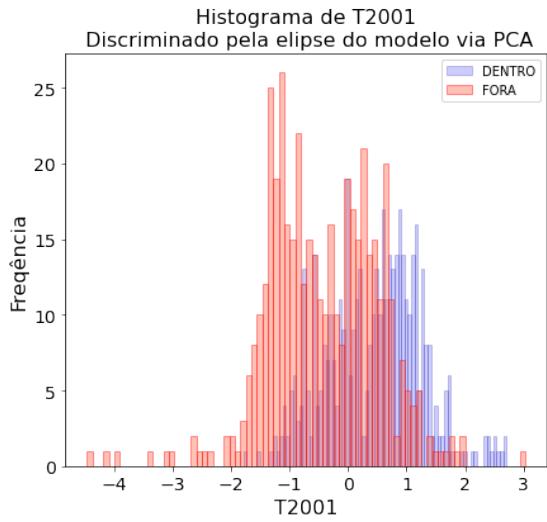
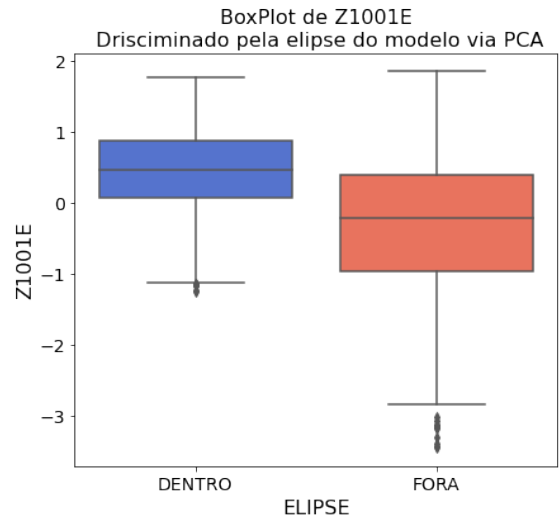
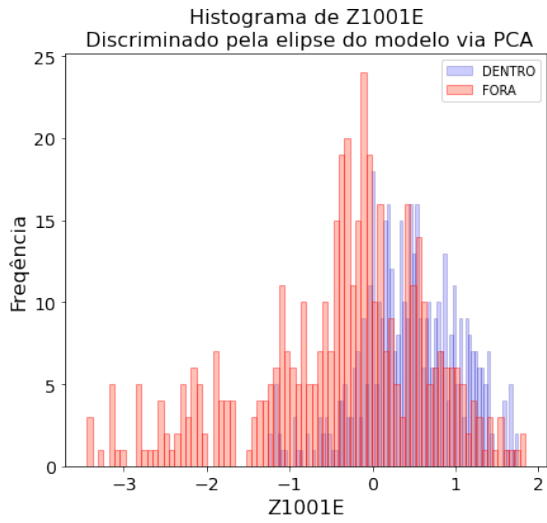
ANEXO A – Histogramas e *Boxplots* das TAGs de *inputs* discriminadas pela Elipse de Confiança definida pelo modelo via PCA

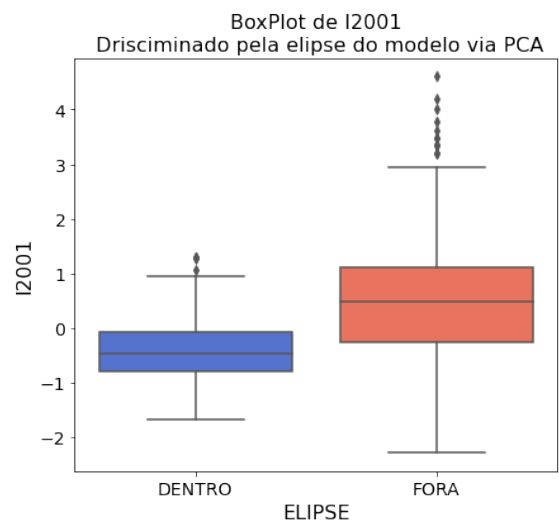
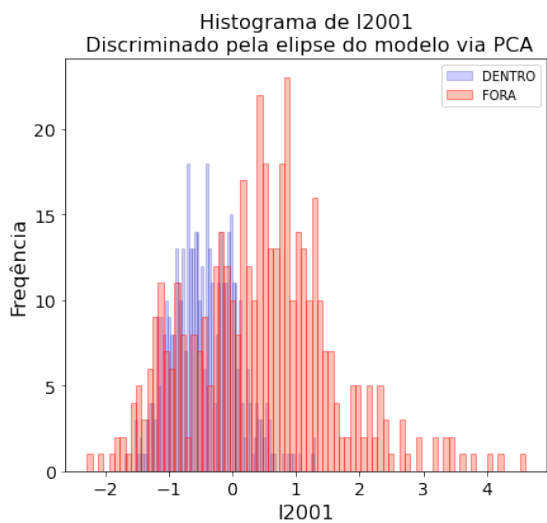
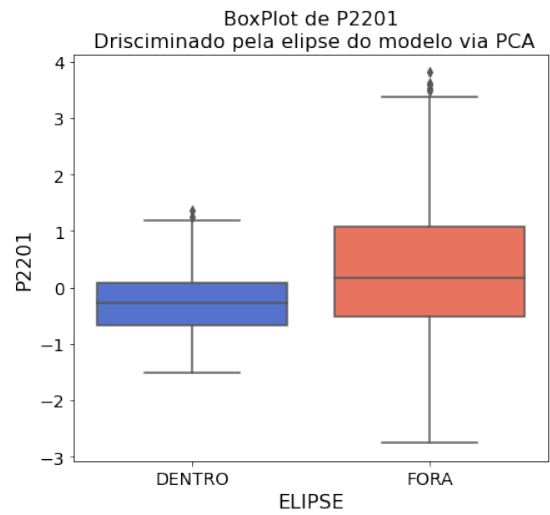
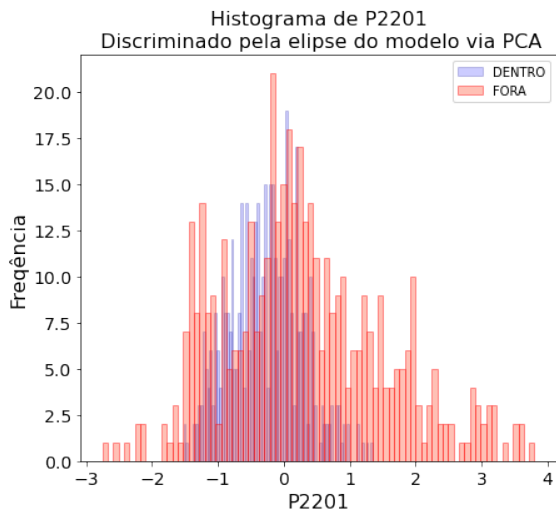
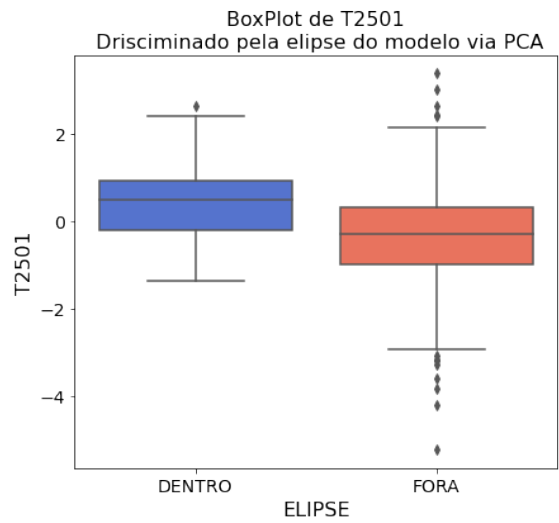
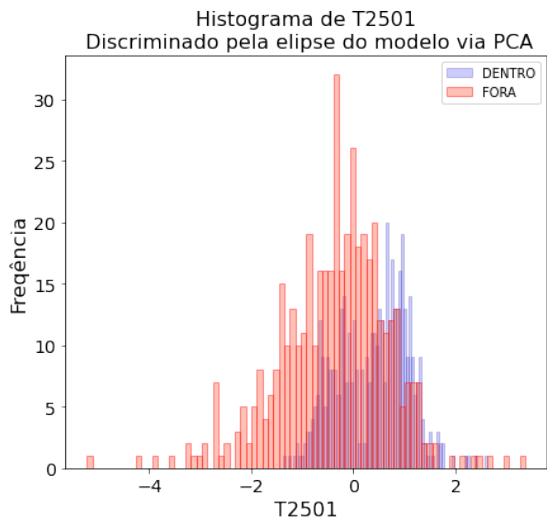


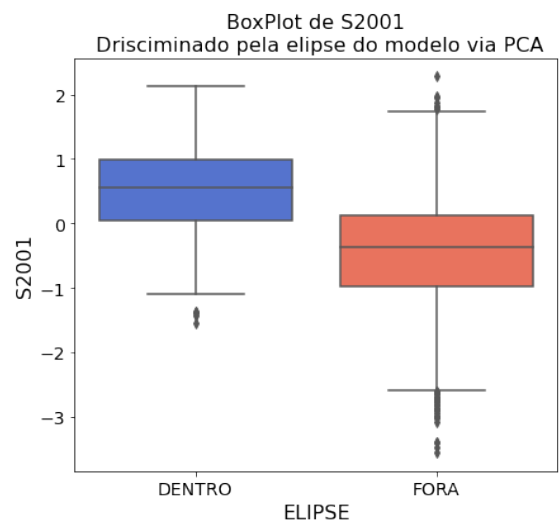
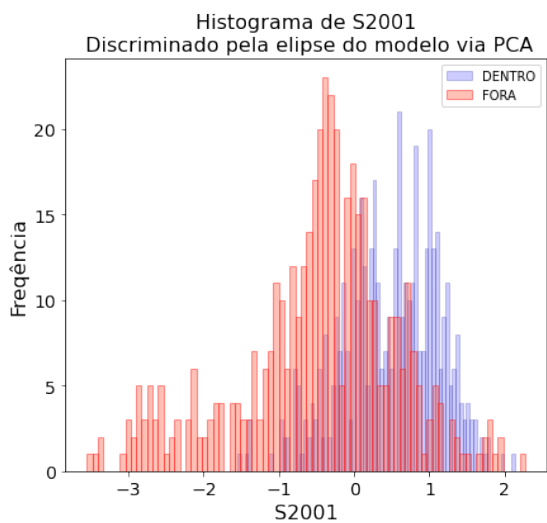
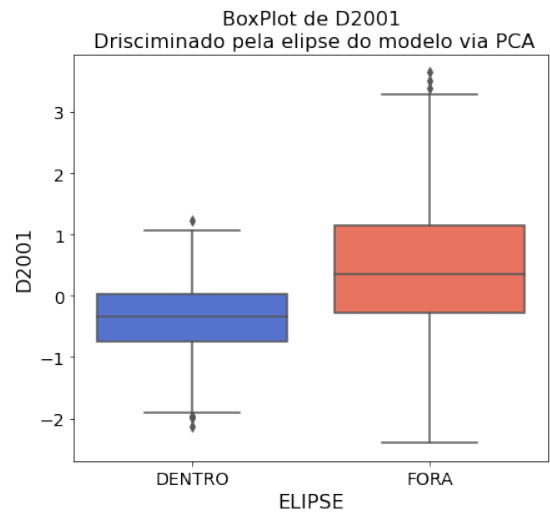
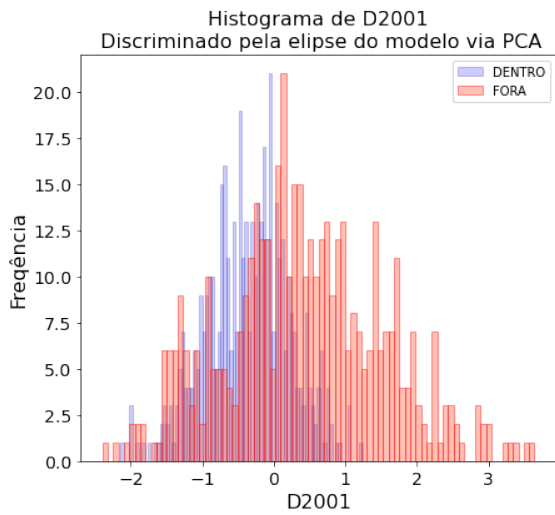
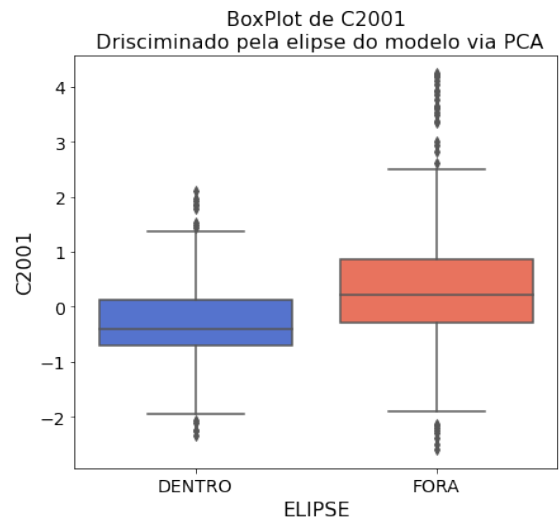
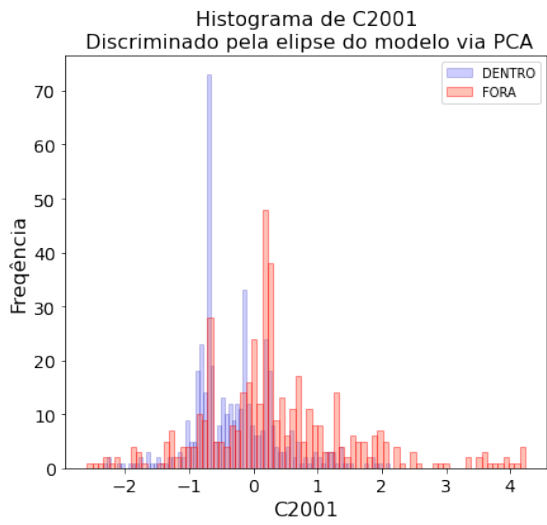












ANEXO B – Histogramas e *Boxplots* das TAGs de *inputs* discriminadas pela Elipse de Confiança definida pelo modelo via FDA

