USO DE TÉCNICAS DE APRENDIZADO DE MÁQUINA PARA PREVISÃO DE FALHAS EM TURBOGERADORES

Gustavo Luís Almeida de Carvalho

Dissertação de Mestrado apresentada ao Programa de Pós-graduação em Engenharia Elétrica, COPPE, da Universidade Federal do Rio de Janeiro, como parte dos requisitos necessários à obtenção do título de Mestre em Engenharia Elétrica.

Orientadores: Sergio Lima Netto
Amaro Azevedo de Lima

Rio de Janeiro
Março de 2018

USO DE TÉCNICAS DE APRENDIZADO DE MÁQUINA PARA PREVISÃO DE FALHAS EM TURBOGERADORES

Gustavo Luís Almeida de Carvalho

DISSERTAÇÃO SUBMETIDA AO CORPO DOCENTE DO INSTITUTO ALBERTO LUIZ COIMBRA DE PÓS-GRADUAÇÃO E PESQUISA DE ENGENHARIA (COPPE) DA UNIVERSIDADE FEDERAL DO RIO DE JANEIRO COMO PARTE DOS REQUISITOS NECESSÁRIOS PARA A OBTENÇÃO DO GRAU DE MESTRE EM CIÊNCIAS EM ENGENHARIA ELÉTRICA.

Examinada por:

_____
Prof. Sergio Lima Netto, Ph.D.

_____
Prof. Amaro Azevedo de Lima, Ph.D.

_____
Prof. Eduardo Antônio Barros da Silva, Ph.D.

_____
Prof. Diego Barreto Haddad, D.Sc.

RIO DE JANEIRO, RJ – BRASIL
MARÇO DE 2018

Carvalho, Gustavo Luís Almeida de

Uso de Técnicas de Aprendizado de Máquina para Previsão de Falhas em Turbogeradores/Gustavo Luís Almeida de Carvalho. – Rio de Janeiro: UFRJ/COPPE, 2018.

XIV, 93 p.: il.; 29, 7cm.

Orientadores: Sergio Lima Netto

Amaro Azevedo de Lima

Dissertação (mestrado) – UFRJ/COPPE/Programa de Engenharia Elétrica, 2018.

Referências Bibliográficas: p. 90 – 93.

1. Aprendizado de Máquina. 2. Manutenção Baseada em Condição. 3. Detecção de Falhas. I. Netto, Sergio Lima *et al.* II. Universidade Federal do Rio de Janeiro, COPPE, Programa de Engenharia Elétrica. III. Título.

Agradecimentos

Agradeço, primeiramente, a Deus e a meus pais pela oportunidade de viver.

Agradeço também a minha família, especialmente a minha mãe, por sempre me incentivar a buscar e oferecer o meu melhor, e a minha esposa, Renata, por sempre me acompanhar com toda sua atenção e amor em todas situações.

Agradeço aos meus orientadores, Professores Sergio Lima Netto e Amaro Azevedo de Lima, por guiar meu aprendizado neste árduo caminho que é o mundo da pesquisa.

Agradeço ao Instituto Alberto Luiz Coimbra de Pós-Graduação e Pesquisa em Engenharia e ao Laboratório de Sinais, Multimídia e Telecomunicações por oferecerem um corpo docente de excelência e ótimas instalações para estudo e pesquisa.

Agradeço, por fim, à Petróleo Brasileiro S.A pela oportunidade de desenvolvimento deste trabalho, acreditando que a melhoria contínua na capacitação de seus funcionários é fundamental para seu crescimento.

Resumo da Dissertação apresentada à COPPE/UFRJ como parte dos requisitos necessários para a obtenção do grau de Mestre em Ciências (M.Sc.)

USO DE TÉCNICAS DE APRENDIZADO DE MÁQUINA PARA PREVISÃO DE FALHAS EM TURBOGERADORES

Gustavo Luís Almeida de Carvalho

Março/2018

Orientadores: Sergio Lima Netto
Amaro Azevedo de Lima

Programa: Engenharia Elétrica

Na indústria, a capacidade de detecção de anomalias nas condições operacionais é de grande interesse. Se identificadas com a antecedência adequada, as intervenções de manutenção podem ser planejadas sob demanda, o que determina um programa de manutenção baseada em condição. Com o aumento da quantidade de dados adquiridos para supervisão e do poder computacional para processamento, o desenvolvimento de técnicas de aprendizado de máquina pode auxiliar na detecção de condições de operação que indiquem necessidade de manutenção. Nesta dissertação, a aplicação destas técnicas é estudada para permitir a identificação de falhas em partidas de turbogeradores. Apresentam-se metodologias para o tratamento dos bancos de dados de operação, para a seleção de variáveis e para o levantamento de características que representem os casos operacionais adequadamente. Classificadores são projetados a partir destes dados e comparados entre si para avaliar a eficácia destes métodos.

Abstract of Dissertation presented to COPPE/UFRJ as a partial fulfillment of the requirements for the degree of Master of Science (M.Sc.)

ON THE USE OF MACHINE LEARNING TO PREDICT FAULTS IN TURBO GENERATORS

Gustavo Luís Almeida de Carvalho

March/2018

Advisors: Sergio Lima Netto
Amaro Azevedo de Lima

Department: Electrical Engineering

In the industry, anomaly detection capability under the operating conditions is of great interest. If identified well in advance, maintenance interventions can be planned on demand, which determines a condition based maintenance. With the increase of the amount of data acquired for supervision and of the computational power for processing, the development of machine learning techniques can aid in the detection of operating conditions that indicate maintenance needs. In this dissertation, such techniques are applied to allow the identification of failures in turbo generators. Methodologies are presented for the treatment of the operation databases, for the selection of variables and for the identification of characteristics that represent the operational cases properly. Classifiers are designed with this data and compared to each other to evaluate the effectiveness of these methods.

Sumário

Lista de Figuras

Lista de Tabelas

Lista de Abreviaturas

CBM	*Condition based maintenance* (manutenção baseada em condição), p. 1
CP	*Componentes principais*, p. 23
DLF	*WHRU inlet / bypass damper linkage failure*, tipo de evento de PCF, p. 5
ETH	*Exhaust average temperature high shutdown*, tipo de evento de PCF, p. 5
FPSO	*Floating production storage & offloading*, p. 2
GDF	*Gas downstream pressure fault shutdown*, tipo de evento de PCF, p. 5
GUF	*Gas upstream pressure fault shutdown*, tipo de evento de PCF, p. 5
IGF	*Ignition failure shutdown*, tipo de evento de PCF, p. 5
KPCA	*Kernel Principal Component Analysis*, p. 37
LUF	*Liquid upstream pressure fault shutdown*, tipo de evento de PCF, p. 5
MSE	*Mean square error* (erro quadrático médio), p. 20
OIF	*Overfuel to ignition failure shutdown*, tipo de evento de PCF, p. 5
PCA	*Principal component analysis* (análise de componentes principais), p. 2
PCF	Partida com falha, p. 9
PFA	*Principal feature analysis*, p. 22

PSF Partida sem falha, p. 9

RLT *Main LO rundown tank fill timeout*, tipo de evento de PCF, p. 5

TG Turbogerador, p. 4

VPE *LVDT VIGV 2/3 position error shutdown*, tipo de evento de PCF, p. 5

Capítulo 1

Introdução

Na indústria, objetiva-se maximizar a produção e minimizar os intervalos de parada funcional. No entanto, os equipamentos utilizados apresentam desgastes ao longo de sua campanha. Para mitigar efeitos negativos decorrentes do uso contínuo, os fabricantes fornecem planos de manutenção que indicam a periodicidade de revisões e de trocas de insumos, que refletem na disponibilidade da planta industrial. Em geral, estes períodos recomendados possuem margens de segurança associadas para a proteção do equipamento.

Não respeitar o plano de manutenção conforme indicado pelos manuais e realizar intervenções em equipamentos somente após a ocorrência de falha representam um risco elevado (*breakdown maintenance* [1]), pois os efeitos colaterais decorrentes da operação fora das condições ideais podem até implicar danos permanentes. Por outro lado, realizar todas as manutenções indicadas pelo fornecedor não exclui a possibilidade de falha do equipamento em prazo menor que aquele estimado pelo fabricante. Novamente, o tempo de parada de operação pode ser longo, a depender da extensão dos danos decorrentes de uma falha não prevista.

Com a constante evolução de tecnologias de detecção e da capacidade de armazenamento e processamento dos dados de operação, surge o conceito de manutenção baseada em condição, também conhecida como *condition based maintenance* (CBM) [1, 2]. Através do acompanhamento dos parâmetros de operação, deseja-se determinar, com a antecedência adequada, as necessidades de intervenções para manutenção. Desta forma, o funcionamento da instalação seria otimizado.

Ao passo que a análise destes dados pode ser realizada exclusivamente pela equipe de operação, várias técnicas de processamento de sinais são estudadas [2–4] e implantadas para facilitar a tarefa de diagnóstico. Estas técnicas são categorizadas como *machine learning*, ou aprendizado de máquina. As tarefas de pré-processamento, geração de novas representações de conjuntos de dados, análise de regras de funcionamento e criação de classificadores ou regressores são exemplos de técnicas de aprendizado de máquina [4] utilizadas para CBM.

Esta dissertação analisa a detecção de falhas em turbogeradores instalados em unidades do tipo *floating production storage & offloading* (FPSO). Conforme [2], estes equipamentos são ativos críticos, com alto custo de manutenção e alta disponibilidade requerida para manter a continuidade operacional. Através do histórico de operação deste tipo de equipamento, percebeu-se que, em geral, os turbogeradores possuem alta confiabilidade operacional. Com isso, constatou-se a possibilidade de redução de custos de manutenção e de aumento de disponibilidade da unidade através da aplicação de CBM nestes equipamentos.

Após realizar uma análise criteriosa das principais falhas que incidem nos turbogeradores, elas foram ranqueadas pela sua frequência e pelo seu custo de manutenção associados. Os tipos de falha priorizados estavam diretamente ligados ao sistema de controle de combustível e implicavam partidas mal sucedidas destes equipamentos. No trabalho proposto em [3], foram desenvolvidos classificadores para a detecção destas falhas a partir de dois bancos de dados de operação da unidade. O primeiro consiste no conjunto de séries temporais de variáveis de processo associadas ao funcionamento do sistema de geração de energia. Os eventos de operação e manutenção anotados dos turbogeradores constituem o segundo banco de dados. Adotou-se como premissa que os dados de operação de 24 horas ininterruptas de funcionamento que antecedem uma falha estão correlacionados com a sua ocorrência.

Mantendo as premissas originais de [3], o foco principal deste trabalho é a obtenção de conjuntos de dados de entrada capazes de melhorar a capacidade de separação dos casos de falha em relação aos casos de operação normal. A partir do uso da análise de componentes principais (PCA) e de técnicas dela derivadas, são propostos métodos de seleção de variáveis de processo e de extração de novas representações destes sinais. Classificadores são projetados a partir destes dados e comparados entre si para avaliar a eficácia destes métodos.

Os bancos de dados originais possuem informações referentes à operação no período de janeiro de 2010 a dezembro de 2012. Esta pesquisa realiza também a ampliação destes bancos de dados, adquirindo dados operacionais de janeiro de 2013 até julho de 2016. Esta extensão permite estudar a variação do perfil operacional dos turbogeradores e avaliar as modelagens de dados propostas. Por fim, os métodos propostos foram aplicados no banco de dados completo para a análise da influência dos novos dados de operação na geração de novos conjuntos de dados de entrada.

1.1 Organização da dissertação

Esta seção apresenta a distribuição dos assuntos através dos capítulos do texto.

No Capítulo 2, a motivação da análise de falhas em turbogeradores é apresentada. Os dados de entrada são descritos e os experimentos realizados em [3] são detalhados

e reproduzidos.

O Capítulo 3, propõe um método de seleção de variáveis de entrada para o processo de treinamento e classificação de falhas em partidas dos equipamentos. No Capítulo 4, esta análise é estendida, buscando representações mais adequadas para o conjunto de dados dos turbogeradores.

O Capítulo 5 apresenta novos experimentos realizados após a aquisição de uma extensão do banco de dados original, que testam a robustez dos classificadores treinados e das variáveis selecionadas nos capítulos anteriores.

Por fim, no Capítulo 6, as conclusões desta dissertação são apresentadas, indicando sugestões de continuação desta pesquisa.

Capítulo 2

Detecção de Falhas em Turbogeradores

2.1 Introdução

Há uma grande variedade de aplicações práticas que usam *machine learning* como alternativa para o processamento de dados. Esta dissertação tem como objetivo aprofundar o estudo realizado por [2, 3] para detecção de falhas em partidas de turbogeradores. Para isso, é necessário o entendimento adequado do trabalho proposto por estes artigos.

Este capítulo apresenta, na seção 2.2, os motivos para a priorização destes equipamentos e descreve os sistemas associados a eles. A seção 2.3 detalha os dados disponíveis para análise e suas principais características. As etapas de pré-processamento, modelagem e seleção dos dados são descritas nas seções 2.4.1 e 2.4.2. Os experimentos realizados em [3] foram analisados, detalhados e reproduzidos na seção 2.4.3. A análise dos resultados obtidos e as oportunidades de refinamento das técnicas apresentadas são mostradas na seção 2.5.

2.2 Motivação

As pesquisas que antecederam e motivaram esta dissertação foram desenvolvidas a partir da análise de quatro turbogeradores (TG) de uma unidade do tipo FPSO. Os turbogeradores são compostos de turbinas aeroderivadas de capacidade nominal de 25 MW, que movimentam um gerador elétrico e podem trabalhar com frequência fixa (modo isócrono) ou variante em função da carga (modo *droop*). A carga requerida para alimentar a FPSO varia tipicamente entre 35∼45 MW. Desta forma, com cada turbogerador arcando com 12∼15 MW em modo isócrono, ao menos 3 máquinas devem operar simultaneamente [2].

Considerando a importância da continuidade operacional destas máquinas e seu alto custo de aquisição e manutenção, o estudo aprofundado destes equipamentos traz ganhos operacionais e, consequentemente, econômicos para FPSOs equipadas com estes turbogeradores.

O artigo [2] apresenta toda a metodologia utilizada para orientar os estudos destas máquinas. Primeiramente, com o auxílio das normas ISO [5–10], os turbogeradores foram separados em subsistemas. Estas normas também definem quais são as principais variáveis de processo que devem ser observadas para acompanhar adequadamente seu funcionamento.

Realizou-se em paralelo uma análise da frequência de ocorrência e dos custos associados a cada evento de manutenção anotado. Como resultado, foi gerada uma lista de priorização de eventos, que é apresentada na Tabela 2.1.

Tabela 2.1: Priorização de Eventos. Fonte: [2]

Evento
FC: OVERFUEL TO IGNITION FAILURE SHUTDOWN (OIF)
FC: GAS DOWNSTREAM PRESSURE FAULT SHUTDOWN (GDF)
FC: GAS UPSTREAM PRESSURE FAULT SHUTDOWN (GUF)
FC: LIQ UPSTREAM PRESSURE FAULT SHUTDOWN (LUF)
FC: IGNITION FAILURE SHUTDOWN (IGF)
GG EXHAUST AVERAGE TEMPERATURE HIGH SHUTDOWN (ETH)
IS-SD: MAIN LO RUNDOWN TANK FILL TIMEOUT (RLT)
WHRU INLET/BYPASS DAMPER LINKAGE FAILURE (DLF)
FC: LVDT VIGV 2/3 POSITION ERROR SHUTDOWN (VPE)

O objetivo dos estudos realizados nos artigos [2, 3] é a criação de ferramentas para auxiliar tomadas de decisão da equipe de operação das FPSOs. Através do processamento dos dados de operação prévios, deseja-se diagnosticar modos de funcionamento que indiquem a necessidade de intervenções para manutenção dos equipamentos. A adoção de ações de manutenção condicionadas pode permitir tanto o aumento do tempo de campanha das FPSOs quanto a antecipação de falhas inesperadas.

2.3 Dados de entrada

Dois conjuntos de dados foram utilizados para a análise. O primeiro é composto de variáveis de processo que são adquiridas e disponibilizadas durante a operação das máquinas. Através delas, os operadores são capazes de acompanhar seu desempenho e tomar ações de acordo com as necessidades da FPSO. Estes dados consistem, tipicamente, em medições de temperatura, vibração, pressão, dentre outras grandezas.

As séries históricas de todas as variáveis disponíveis são coletadas dos servidores de dados históricos da FPSO. A unidade estudada possui quatro turbogeradores,

nomeados de TGA, TGB, TGC e TGD.

Foi fornecido um total de 442 variáveis (*tags*) distintas registradas no banco de dados. Verificou-se que somente 86 *tags*, listadas na Tabela 2.2, eram comuns a todos os quatro turbogeradores. Desta forma, as *tags* remanescentes não foram utilizadas.

O período de aquisição de dados selecionado foi de fevereiro de 2010 a dezembro de 2012, pois foi o período em que todas as 86 *tags* possuíam leituras válidas. As leituras fora deste período não foram analisadas.

Tabela 2.2: Variáveis de processo fornecidas.

Índice	***Tag* TGA/TGB/TGC/TGD**	**Descrição**
1	FIT-001 A/B/C/D	Vazão de gás combustível
2	TI-001 A/B/C/D	Temperatura do tanque de óleo sintético
3	TI-002 A/B/C/D	Temperatura do *header* de óleo mineral
4	TIT-001 A/B/C/D	Saída de água quente do WHRU
5	TIT-002 A/B/C/D	Temperatura do WHRU
6	TT-001 A/B/C/D	Temperatura ambiente
7	TT-002 A/B/C/D	Temperatura do gás combustível
8	TT-003 A/B/C/D	Temperatura do *manifold* de gás combustível
9	VE-001 A/B/C/D	Vibração na entrada do GG
10	VE-002 A/B/C/D	Vibração no centro do GG
11	VE-003 A/B/C/D	Vibração na turbina do GG
12	VE-004 A/B/C/D	Vibração PT disc. end. X
13	VE-005 A/B/C/D	Vib. PT disc. end. Y
14	VE-006 A/B/C/D	Vib. PT acopl. X
15	VE-007 A/B/C/D	Vib. PT acopl. Y
16	ZE-001 A/B/C/D	Vib. PT AXIAL
17	ZE-002 A/B/C/D	Vib. PT AXIAL
18	ZE-003 A/B/C/D	*Gearbox* LSS Axial
19	ZE-004 A/B/C/D	*Gearbox* HSS Axial
20∼36	TI-003-01∼17 A/B/C/D	Termopares 01 a 17 do perfil da exaustão
37	TI-004 A/B/C/D	Gerador temp. enrolamento L2-3
38	TI-005 A/B/C/D	Gerador temp. enrolamento L2-2
39	TI-006 A/B/C/D	Gerador temp. enrolamento L2-1
40	PDT-001 A/B/C/D	Pressão diferencial - Entrada GG
41	PDT-002 A/B/C/D	Pressão diferencial - Filtro de gás combustível
42	PDI-001 A/B/C/D	Pressão diferencial - Filltro de ar
43	PI-001 A/B/C/D	Pressão *header* óleo mineral
44	PIT-001 A/B/C/D	Pressão
45	PT-001 A/B/C/D	Pressão P1
46	PT-002 A/B/C/D	Exaustão GG
47	PT-003 A/B/C/D	Entrada WHRU
48	PT-003 A/B/C/D	Gás Comb.
49	ST-001 A/B/C/D	Rot. Com. / Turb. Baixa
50	ST-002 A/B/C/D	Rotação Motor de Arranque

51	ST-003 A/B/C/D	Rot. Com. / Turb. Alta
52	ST-004 A/B/C/D	Rotação PT
53	TE-001 A/B/C/D	Exaustão GG
54	IT-001 A/B/C/D	Corrente de excitação do campo
55	ET-001 A/B/C/D	Voltagem ab
56	ET-001 A/B/C/D	Voltagem bc
57	ET-001 A/B/C/D	Voltagem ca
58	JT-001 A/B/C/D	Potência Reativa
59	JT-002 A/B/C/D	*Active Power*
60	JY-001 A/B/C/D	Energia Real
61	JQ-001 A/B/C/D	*Power Factor*
62	ST-005 A/B/C/D	*Frequency*
63	GT-001 A/B/C/D	*Fuel Type*
64	KT-001 A/B/C/D	Horas Carga Gás
65	KT-002 A/B/C/D	Horas Carga Diesel
66	QT-001 A/B/C/D	Partidas Gás
67	QT-002 A/B/C/D	Partidas Diesel
68	IT-002 A/B/C/D	Corrente
69	TI-007 A/B/C/D	Gerador temp. enrolamento L1-3
70	TI-008 A/B/C/D	Exaustão GG
71	TI-009 A/B/C/D	PT Temperatura Mancal NDE
72	TI-010 A/B/C/D	PT Temperatura Mancal de Escora
73	TI-011 A/B/C/D	PT Temperatura mancal DE
74	TI-012 A/B/C/D	*Gearbox* LSS Temperatura Mancal de Escora
75	TI-013 A/B/C/D	*Gearbox* LSS Temperatura mancal DE
76	TI-014 A/B/C/D	*Gearbox* LSS Temperatura mancal NDE
77	TI-015 A/B/C/D	*Gearbox* HSS Temperatura mancal NDE
78	TI-016 A/B/C/D	Gerador ar de resfriamento (frio) DE
79	TI-017 A/B/C/D	Gerador ar de resfriamento (quente) DE
80	TI-018 A/B/C/D	Gerador ar de resfriamento (quente) NDE
81	TI-019 A/B/C/D	Gerador ar de resfriamento (frio) NDE
82	TI-020 A/B/C/D	Gerador Temperatura mancal DE
83	TI-021 A/B/C/D	Gerador temp. enrolamento L1-1
84	TI-022 A/B/C/D	Gerador temp. enrolamento L1-2
85	TI-023 A/B/C/D	Gerador temp. enrolamento L3-1
86	TI-024 A/B/C/D	Gerador temp. enrolamento L3-2

A segunda fonte de informação é formada pelo registro de eventos relacionados à operação e à manutenção dos turbogeradores. Os eventos são registrados manualmente em um *software* proprietário. O *software* é capaz de exportar seus dados para planilhas de acompanhamento. Para cada entrada desta tabela, os seguintes parâmetros são fornecidos:

- Máquina: identifica o turbogerador afetado;

- Data / hora do início do evento;
- Data / hora do final do evento;
- Total de horas do evento;
- Identificação e descrição padrões do evento: apresentadas conforme programação padrão do fabricante;
- Complemento da descrição do evento: apresenta detalhes adicionais da operação da unidade durante o evento;
- Ações decorrentes do evento: indica providências decorrentes da parada;
- Tipo de parada: indica se a parada foi programada ou decorrente de falhas;
- Subsistema afetado: indica qual subsistema foi afetado na parada;
- Ocorrência de falha na partida: indica se houve falha na partida após parada;
- Horímetro;
- Número de partidas;
- Partida em *cooldown*.

A associação destes conjuntos de dados permite criar subconjuntos contendo somente as variáveis mais relevantes nos períodos de ocorrência de cada evento anotado. Na seção 2.4 são descritas todas as etapas de processamento destas informações para a execução dos algoritmos de identificação propostos no artigo [3].

2.4 Processamento e análise

O tratamento dos conjuntos de dados é separado em 3 etapas. A primeira etapa consiste no pré-processamento, onde todas as inconsistências e valores fora da faixa de medição esperados são excluídos. Já a segunda etapa é responsável pela seleção das séries temporais, dos intervalos de interesse e dos parâmetros que melhor descrevem os eventos observados. Por fim, a terceira etapa reúne todos os experimentos que foram desenvolvidos em busca da melhor capacidade de identificação dos modos de funcionamento dos turbogeradores.

2.4.1 Pré-processamento

• Séries de dados históricos

O *software* historiador utilizado é o PI System, da OSIsoft. A taxa de amostragem destes dados é de uma amostra por minuto. Com o intuito de otimizar a ocupação de espaço de armazenamento, toda sequência de amostras que possuem uma relação linear é excluída pelo PI, guardando-se somente as amostras inicial e final deste conjunto de pontos. Para que todas as séries temporais analisadas possuam taxa de amostragem constante, os trechos com linearidade suprimidos pelo historiador foram recuperados através de interpolação linear.

Dentro do sistema de aquisição de dados, falhas nos sensores, nos cartões de entrada, no processamento de dados ou no *software* historiador podem resultar em valores medidos fora da faixa de medição. Para cada série temporal, através da aplicação de um filtro de limiar, foram identificadas todas as amostras fora da faixa de medição. Seus valores foram substituídos pela média de suas amostras vizinhas.

• Eventos anotados

Nos experimentos realizados, todos os eventos de partida sem falhas (PSF) foram utilizados. Nos casos de partida com falha (PCF), foram escolhidas somente falhas do tipo OIF, GDF e GUF, pois, no período de aquisição estudado, estes tipos de falha apresentavam maior custo de reparo e um número mais significativo de ocorrências. Estas falhas são, de acordo com [3], associadas ao mau funcionamento da válvula de controle de injeção de combustível. Outras falhas não foram tratadas em [3].

Dentre os registros anotados, notaram-se alguns casos de ocorrência de PCFs consecutivas em curtos intervalos de tempo. Na Figura 2.1, a evolução de um trecho da série temporal de um sensor de temperatura evidencia, através de picos de curta duração, a ocorrência de partidas com falha causadas por GUFs de forma consecutiva. Pode-se concluir que a causa destas falhas é comum e que a manutenção da primeira falha ainda não foi realizada com êxito. Desta forma, somente a primeira parada anotada é considerada.

2.4.2 Seleção de dados

• Premissas

Após o pré-processamento das informações, foram selecionados somente os casos de partidas que foram precedidos de, ao menos, 24 horas de operação contínua. Desta forma, para o período estudado, foram selecionados 193 PSFs e 33 PCFs. A distribuição das PSFs e das PCFs entre os turbogeradores é mostrada na tabela 2.3.

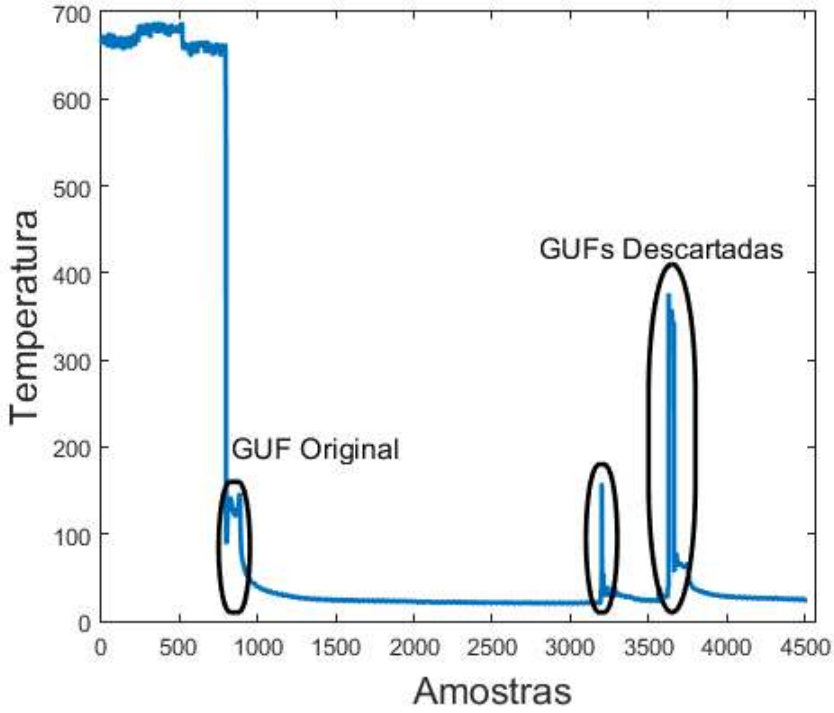


Figura 2.1: Quando há a ocorrência de partidas com falha consecutivas (GUFs neste caso), identificadas por picos de curta duração, considera-se somente a primeira delas.

Tabela 2.3: Distribuição dos diferentes tipos de parada por turbogerador.

	PSF	**PCF-OIF**	**PCF-GDF**	**PCF-GUF**
TGA	49	2	3	1
TGB	47	1	1	7
TGC	48	10	0	0
TGD	49	5	0	3

- **Escolha de variáveis**

Um conjunto de 22 sinais foi utilizado nos experimentos. Conforme apresentado no estudo do artigo [2], verificou-se que o perfil de temperatura no anel de exaustão da turbina pode ser correlacionado com as PCFs. Este perfil de temperatura é obtido através da monitoração de 17 termopares distribuídos ao longo deste anel. Adicionalmente, foram consideradas medições de outras 5 variáveis relacionadas ao sistema de combustão. A Tabela 2.4 apresenta a lista destas variáveis.

Ressalta-se que não houve nenhuma metodologia adicional para escolha ou exclusão de variáveis apresentada em [3].

2.4.3 Experimentos desenvolvidos

O grande desbalanceamento no número de PSFs e PCFs representa uma dificuldade para o treinamento de classificadores. A solução adotada utiliza o algoritmo de

Tabela 2.4: Variáveis de processo selecionadas para análise.

Índice	***Tag* TGA/TGB/TGC/TGD**	**Descrição**
1∼17	TI-003-01∼17 A/B/C/D	Termopares 01 a 17 do perfil da exaustão
18	FIT-001 A/B/C/D	Vazão de Gás Combustível
19	PT-002 A/B/C/D	Pressão na exaustão
20	PDT-001 A/B/C/D	Dif. de pressão na entrada do GG
21	TE-001 A/B/C/D	Temperatura na exaustão
22	TI-008 A/B/C/D	Temperatura na exaustão

classificação RUSBoost [11], que foi desenvolvido para lidar com conjuntos de dados desbalanceados. Para verificar o desempenho dos classificadores resultantes foi realizada a divisão dos subconjuntos de dados em seis grupos distintos. Em cada iteração do experimento, 3 grupos são utilizados para o treinamento. O classificador resultante é testado com os 3 grupos restantes. Portanto, cada experimento possui 20 realizações, que correspondem ao número de combinações distintas possíveis destes 6 grupos. Este método é conhecido como validação cruzada (*cross validation*) e é indicado para testes de classificadores quando há poucos casos disponíveis para análise [4, 12].

O treinamento dos classificadores utilizando o RUSBoost foi desenvolvido através da função *fitensemble* do *software* Matlab versão R2015a. Os seguintes parâmetros de configuração foram utilizados:

- Método (parâmetro *Method*): 'RUSBoost';
- Número de ciclos de aprendizagem (parâmetro *NLearn*): 200;
- Tipo de classificador (parâmetro *Learners*): árvores de decisão (objeto *ClassificationTree* do Matlab);
- Taxa de aprendizado (parâmetro *LearnRate*): 0,1.

• Critérios de agrupamento

Para garantir que haja homogeneidade e representatividade adequadas em todos os grupos, os seguintes passos foram adotados:

- Ordenação dos conjuntos de PCF e PSF por equipamento afetado;
- Ordenação por data de ocorrência para cada tipo de parada;
- Composição do grupo através da retirada intercalada de uma unidade de cada um dos conjuntos ordenados até o esgotamento dos mesmos.

O processo de ordenação dos casos de PCF e PSF é ilustrado na Figura 2.2.

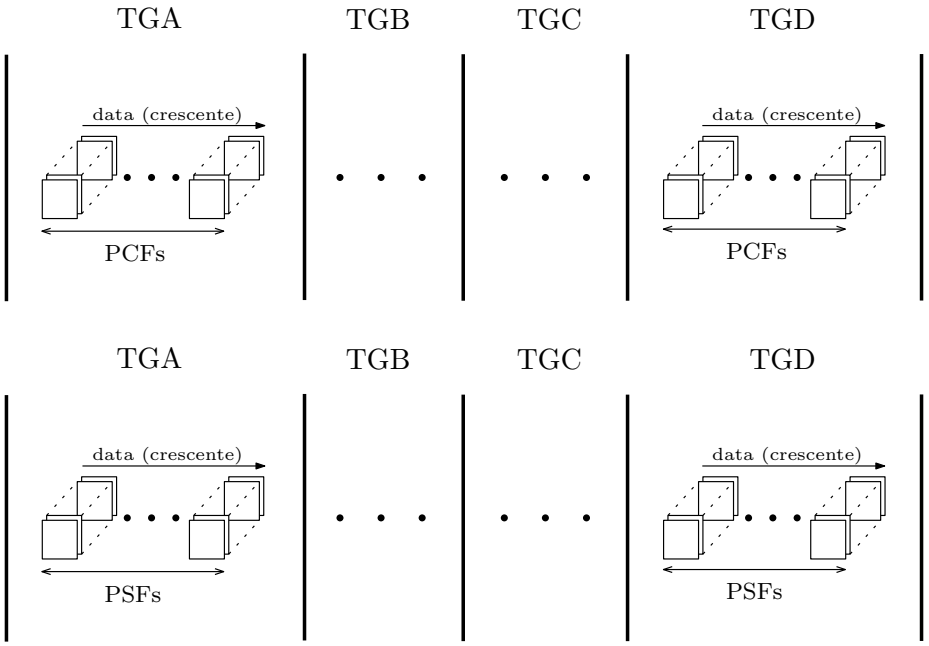


Figura 2.2: A ordenação dos casos de PCF e PSF é realizada através do agrupamento por máquinas e por data de ocorrência.

• Criação do vetor de entrada

Todas as amostras das 22 variáveis selecionadas são utilizadas pelo classificador. Logo, cada caso de operação é representado por um vetor de 31680 amostras (1440 medições, equivalentes a 24 horas de operação contínua, por variável). Dentre as diferentes formas possíveis de combinação destes pontos para a formação do vetor resultante, a organização foi realizada pela concatenação das amostras por grupo de variáveis (a ordem das variáveis coincide com a lista da Tabela 2.4). Ou seja, o vetor inicia com as amostras de índice 1 das 22 variáveis e termina com as amostras de índice 1440 destas.

• Experimento 1

O primeiro experimento consistiu em dividir todas as 193 PSFs e 33 PCFs em 6 grupos. Desta forma foram gerados 6 grupos com as seguintes composições:

- 1 grupo com 33 PSFs e 6 PCFs;
- 2 grupos com 32 PSFs e 6 PCFs;
- 3 grupos com 32 PSFs e 5 PCFs.

A distribuição das paradas entre os grupos é mostrada na Tabela 2.5.

Através da execução de cinco repetições consecutivas do experimento, devido à escolha de subconjuntos de PSFs de forma aleatória pelo RUSBoost, foram obtidos

Tabela 2.5: Distribuição das Paradas por Grupos - Experimento 1.

	G1	G2	G3	G4	G5	G6
TGA-PSF	8	9	8	8	8	8
TGA-OIF	0	0	1	0	1	0
TGA-GDF	1	1	0	1	0	0
TGA-GUF	0	0	0	0	0	1
TGB-PSF	7	8	8	8	8	8
TGB-OIF	0	0	0	1	0	0
TGB-GDF	0	1	0	0	0	0
TGB-GUF	2	1	1	1	1	1
TGC-PSF	8	8	8	8	8	8
TGC-OIF	1	2	2	1	2	2
TGD-PSF	9	8	8	8	8	8
TGD-OIF	2	1	1	1	0	0
TGD-GUF	0	0	0	1	1	1

resultados similares ao artigo [3]. Na fase de treinamento, a taxa de acerto médio para PCFs foi de 100% e de 86,36% para PSFs. A Figura 2.3(a) mostra a variação destas taxas para cada uma das combinações de grupos.

Na fase de testes, a taxa de acerto médio para PCFs foi de 53,46% e de 65,44% para PSFs. As Figuras 2.4(a) e 2.5(a) mostram, respectivamente, a variação das taxas de acerto médio para cada combinação e a média das taxas de acerto médio das PCFs e PSFs acompanhadas de seus desvios-padrão.

- **Experimento 2**

Percebeu-se que na fase de treinamento havia uma maior taxa de erros de classificação para os casos de PSF. Para tentar amenizar os efeitos nos resultados nos testes dos classificadores, todos os casos de PSF que foram classificados como PCF durante a fase de treinamento no Experimento 1 foram descartados.

Com as 70 PSFs restantes, a distribuição das paradas entre os turbogeradores é mostrada na Tabela 2.6.

Tabela 2.6: Distribuição das Paradas por Turbogerador - Experimento 2.

	PSF	PCF-OIF	PCF-GDF	PCF-GUF
TGA	21	2	3	1
TGB	18	1	1	7
TGC	16	10	0	0
TGD	15	5	0	3

Com isso, as novas composições dos grupos são:

- 3 grupos com 12 PSFs e 6 PCFs;
- 1 grupo com 12 PSFs e 5 PCFs;
- 2 grupos com 11 PSFs e 5 PCFs;

A distribuição das paradas entre os grupos é mostrada na Tabela 2.7.

Tabela 2.7: Distribuição das Paradas por Grupos - Experimento 2.

	G1	**G2**	**G3**	**G4**	**G5**	**G6**
TGA-PSF	4	4	3	4	3	3
TGA-OIF	0	0	1	0	1	0
TGA-GDF	1	1	0	1	0	0
TGA-GUF	0	0	0	0	0	1
TGB-PSF	3	3	3	3	3	3
TGB-OIF	0	0	0	1	0	0
TGB-GDF	0	1	0	0	0	0
TGB-GUF	2	1	1	1	1	1
TGC-PSF	2	3	3	2	3	3
TGC-OIF	1	2	2	1	2	2
TGD-PSF	3	2	3	3	2	2
TGD-OIF	2	1	1	1	0	0
TGD-GUF	0	0	0	1	1	1

Através da execução de cinco repetições consecutivas deste experimento conforme proposto, foram obtidos resultados similares ao artigo [3] novamente. Na fase de treinamento, a taxa de acerto médio para PCFs foi de 100% e de 99,72% para PSFs. A Figura 2.3(b) mostra a variação destas taxas para cada uma das combinações de grupos.

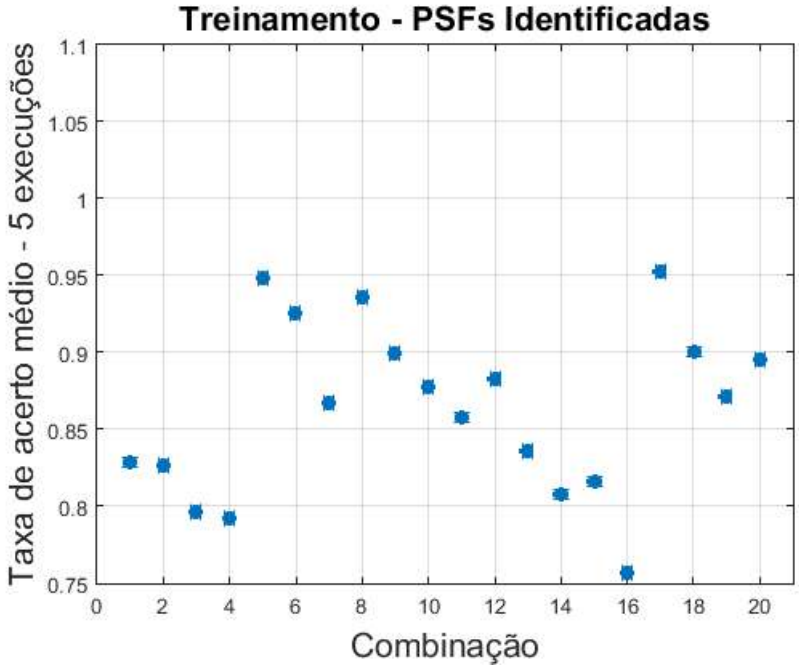
Na fase de testes, a taxa de acerto médio para PCFs foi de 75,63% e de 84,24% para PSFs. As Figuras 2.4(b) e 2.5(b) mostram, respectivamente, a variação das taxas de acerto médio para cada combinação e a média das taxas de acerto médio das PCFs e PSFs acompanhadas de seus desvios-padrão.

2.5 Conclusão

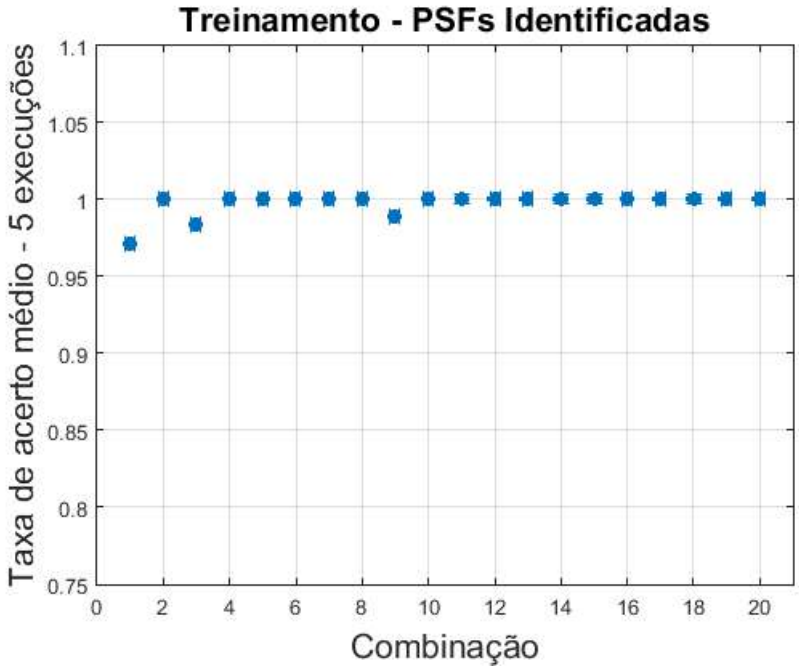
Este capítulo apresentou as técnicas desenvolvidas para detecção de falhas em partidas de turbogeradores, conforme [2, 3]. Foram detalhadas, na seção 2.3, as bases de dados, que contam tanto com séries históricas de variáveis de processo como com dados anotados de manutenção. Os passos necessários para o tratamento adequado destas informações é detalhado na seção 2.4.1. No Capítulo 5, estas técnicas de pré-processamento de dados são utilizadas para extensão dos bancos de dados existentes.

Os experimentos desenvolvidos em [3] foram detalhados e reproduzidos nas seções 2.4.2 e 2.4.3. Destaca-se que foram escolhidas 22 séries temporais de um total de 86 séries através do conhecimento da equipe operacional. No entanto, não foi apresentado estudo que evidencie que as séries restantes não possuam informações relevantes a serem analisadas. Os Capítulos 3 e 4 analisam métodos para determinação de conjuntos de dados que proporcionem melhor capacidade de separação dos conjuntos de PCFs e PSFs.

Conforme evidenciado por [3], há um conjunto de PSFs que apresenta características diferenciadas e, com isso, prejudicavam, já na fase de treinamento, a capacidade de identificação dos classificadores desenvolvidos. A semelhança destes casos com as PCFs estudadas pode indicar estados de operação que precedem falhas nos equipamentos. Desta forma, no Experimento 2, utilizou-se somente um subconjunto de 70 PSFs. Nos Capítulos 3 e 4, para o desenvolvimento de novos classificadores, somente este subconjunto de PSFs é utilizado para análise.

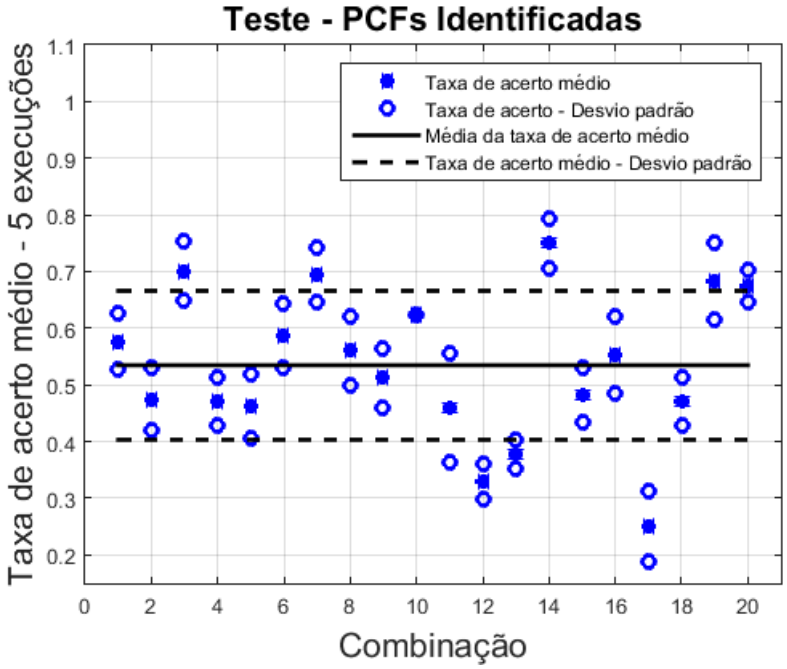


(a) Experimento 1

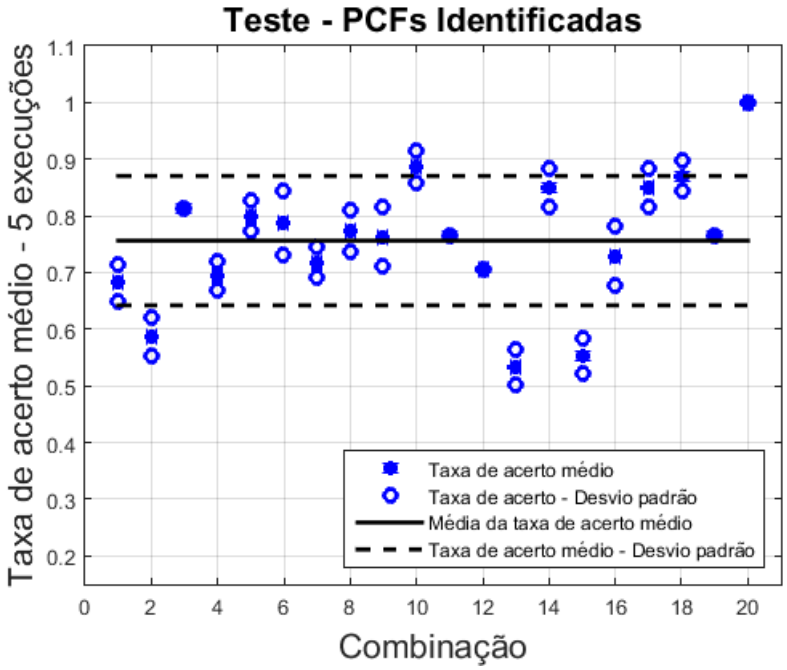


(b) Experimento 2

Figura 2.3: Taxa de acertos para PSFs no treinamento: após a redução do número de PSFs analisadas, desconsiderando PSFs identificadas incorretamente no Experimento 1 (cuja média da taxa de acerto médio obtida foi 86,36%), a média da taxa de acerto médio na etapa de treinamento do Experimento 2 foi elevada para 99,72%.

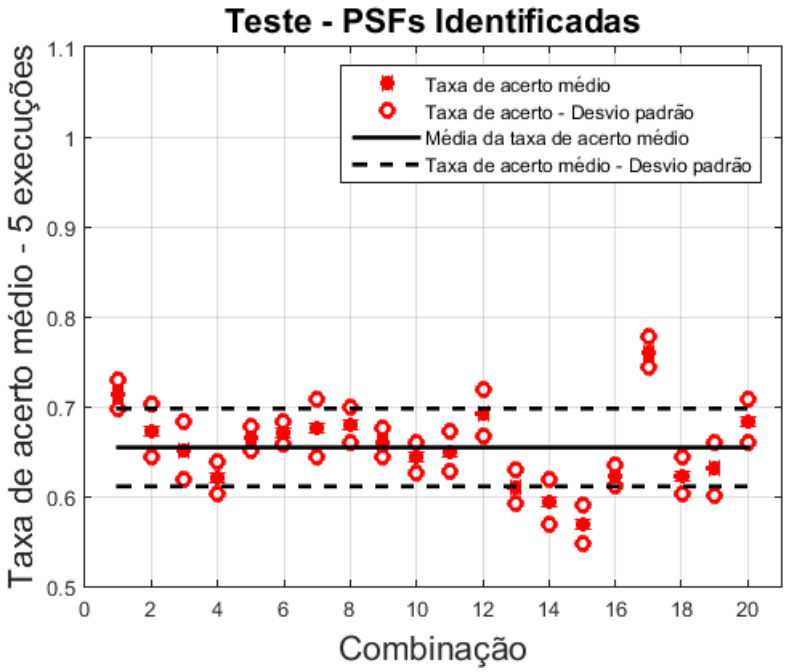


(a) Experimento 1

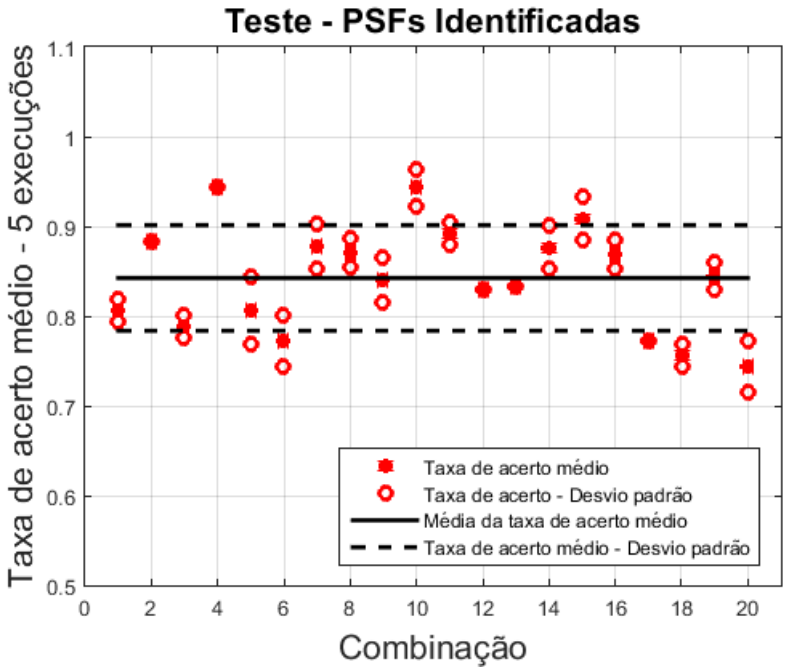


(b) Experimento 2

Figura 2.4: Taxa de acertos para PCFs nos testes: após a redução do número de PSFs analisadas, desconsiderando PSFs identificadas incorretamente no Experimento 1 (cuja média das taxas de acerto médio obtida foi 53,46%), a média das taxas de acerto médio na etapa de teste do Experimento 2 foi elevada para 75,63%.



(a) Experimento 1



(b) Experimento 2

Figura 2.5: Taxa de acertos para PSFs nos testes: após a redução do número de PSFs analisadas, desconsiderando PSFs identificadas incorretamente no Experimento 1 (cuja taxa de acertos média obtida foi 65,44%), a taxa de acertos média na etapa de teste do Experimento 2 foi elevada para 84,24%.

Capítulo 3

Escolha de Variáveis

3.1 Introdução

Nos experimentos apresentados no Capítulo 2, os classificadores desenvolvidos utilizavam, como entrada, períodos de um dia de operação contínua de séries temporais de 22 variáveis monitoradas dos turbogeradores. Sabe-se que a escolha dos sinais foi justificada pela experiência e conhecimento acumulados de equipes de operação.

Contudo, considerando os resultados obtidos, não foram apresentadas evidências de que este conjunto de dados representa a melhor alternativa para o problema. Se a quantidade de variáveis puder ser reduzida, obtendo resultados similares ou melhores, o tempo de treinamento para a geração dos classificadores reduziria. Além disso, dentre as 64 variáveis não selecionadas, podem existir sinais que possuam informações relevantes que não foram mapeadas dentro do histórico de operação destes equipamentos.

A busca do conjunto ótimo de variáveis através da avaliação de todas as combinações possíveis dos sinais disponíveis não é adequada. O tempo necessário para testar todas estas combinações cresce demasiadamente conforme o aumento de informações monitoradas. Ademais, considerando o desgaste natural decorrente do uso do equipamento, a solução ideal pode variar ao longo da vida útil dos turbogeradores. Desta forma, é desejável a utilização de técnicas que permitam estimar de forma não supervisionada ou semisupervisionada as informações de maior relevância para otimizar o desempenho dos algoritmos de classificação.

De acordo com [13–15], a utilização da análise de componentes principais (*Principal Component Analysis* - PCA) é aplicável para a seleção de variáveis e para a criação de conjuntos de dados reduzidos, mantendo boa representatividade em comparação com o conjunto completo. A seção 3.2 apresenta os principais resultados da teoria apresentada em [16, 17]. Destes resultados, derivam as técnicas que são descritas na seção 3.3. Na seção 3.4, é mostrado como a aplicação destas técnicas foi

realizada nos dados monitorados dos turbogeradores e seus resultados. Por fim, a seção 3.5 sintetiza os estudos realizados e indica as principais conclusões das análises propostas.

3.2 PCA - Análise de Componentes Principais

O PCA, também conhecido como Transformada de Karhunen-Loève (KL) ou Transformada de Hotelling [17], é uma técnica que, valendo-se das estatísticas de seus sinais de entrada, gera um conjunto de dados descorrelacionados entre si.

Considerando um vetor coluna $\mathbf{x}$, em que cada elemento representa um sinal distinto de um conjunto de medições proposto, deseja-se aplicar uma matriz de transformação $\mathbf{A}$ que gere um novo vetor $\mathbf{y}$ que possua elementos descorrelacionados, ou seja, dados dois instantes a e b, $E[\mathbf{y}(a)\mathbf{y}(b)] = 0$ quando $a \neq b$:

$$\mathbf{y} = \mathbf{A}^T\mathbf{x}. \tag{3.1}$$

A matriz $\mathbf{R}_x$ de correlação de $\mathbf{x}$ é definida por $E[\mathbf{x}\mathbf{x}^T]$. Utilizando a equação (3.1), a matriz de correlação de $\mathbf{y}$ é definida por

$$\mathbf{R}_y = E[\mathbf{A}^T\mathbf{x}\mathbf{x}^T\mathbf{A}] = \mathbf{A}^T\mathbf{R}_x\mathbf{A}. \tag{3.2}$$

Como $\mathbf{R}_x$ é simétrica, seus autovetores associados a autovalores distintos são ortogonais. Se cada coluna de $\mathbf{A}$ for constituída pelo conjunto de autovetores ortonormais de $\mathbf{R}_x$, $\mathbf{R}_y$ será diagonal, com valores de sua diagonal principal iguais aos autovalores λ_x de $\mathbf{R}_x$. Esta matriz é definida como $\mathbf{\Lambda}$. Os vetores são ordenados de acordo com a ordem decrescente dos autovalores λ_x. Desta forma, os primeiros componentes principais são aquelas que concentram a maior energia.

Uma das propriedades mais importantes do PCA é a representação de $\mathbf{x}$ pelos seus k primeiros componentes principais consistir no subconjunto de k variáveis que apresentam o menor erro quadrático médio (*mean square error* - MSE) para reconstrução linear do conjunto original de dados [17].

Se a média dos sinais de $\mathbf{x}$ for igual a zero, a média de $\mathbf{y}$ também o será e a covariância $\mathbf{\Sigma}_x$ será igual a $\mathbf{R}_\mathbf{x}$. Desta forma, a matriz de covariância também pode ser utilizada:

$$E[\mathbf{x}] = \mathbf{0} \Rightarrow \mathbf{R}_y = \mathbf{\Sigma}_y = \mathbf{A}^T\mathbf{R}_x\mathbf{A} = \mathbf{A}^T\mathbf{\Sigma}_x\mathbf{A} = \mathbf{\Lambda}. \tag{3.3}$$

3.2.1 Aproximação da estatística

Conforme discutido em [16, 17], quando a estatística do sinal não é conhecida, o PCA pode ser aplicado através de aproximações da mesma. Utilizando n conjuntos de amostras de $\mathbf{x}$, calculam-se a aproximação amostral da média, $\bar{\mathbf{x}}$, e a aproximação amostral da covariância, $\mathbf{S_x}$, através das equações

$$\bar{\mathbf{x}} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i, \tag{3.4}$$

$$\mathbf{S_x} = \frac{1}{n-1}\sum_{i=1}^{n}(\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^T. \tag{3.5}$$

Através destas aproximações, a equação (3.3) pode ser aproximada por

$$\bar{\mathbf{x}} = \mathbf{0} \Rightarrow \mathbf{R}_y \approx \mathbf{S}_y = \mathbf{A}^T\mathbf{S}_x\mathbf{A} = \mathbf{\Lambda}. \tag{3.6}$$

Ao passo que um subconjunto de componentes principais obtido pela estatística exata resulta no menor MSE da estimativa linear dos dados originais, um subconjunto de componentes principais obtido através das estatísticas aproximadas resulta na menor distância euclidiana possível em relação aos dados observados.

3.2.2 Dados com diferentes escalas

Segundo discutido em detalhes em [16], o PCA é sensível à escala de medição dos dados observados. Se uma das características possuir uma variância muito maior que as outras devido à existência de escalas de medição diferentes, os componentes principais irão priorizar esta variável em detrimento às outras e mascarar relações de dependência. Uma das alternativas para remediar este problema, ao invés de utilizar a correlação ou a covariância dos sinais originais, consiste na normalização dos vetores $\mathbf{x}$ de entrada pelos seus desvios-padrão. Desta forma, ao realizar a correlação entre os sinais normalizados, obtém-se a matriz dos fatores de correlação ρ. Considerando que os vetores $\mathbf{x}$ possuem média zero, cada elemento desta matriz ρ_{ij} é calculado a partir de

$$\rho_{ij} = \frac{E[\mathbf{x}_i\mathbf{x}_j]}{\sqrt{E[\mathbf{x}_i^2]E[\mathbf{x}_j^2]}}. \tag{3.7}$$

3.2.3 Seleção de componentes principais

Como o PCA possui a capacidade de reter grande parte da variabilidade do conjunto em uma pequena quantidade de componentes principais, resta definir um critério para escolher quantos componentes principais devem ser mantidos.

Considerando q componentes principais selecionados, a variabilidade retida V_{retida} por este subconjunto, definida em [14], é calculada através da razão entre o somatório dos q autovalores correspondentes aos componentes principais selecionados e o somatório de todos autovalores λ_x, isto é,

$$V_{\text{retida}} = \frac{\sum_{i=1}^{q} \lambda_i}{\sum_{i=1}^{n} \lambda_i}. \tag{3.8}$$

Ao estipular um valor mínimo de variabilidade, é possível definir o tamanho do subconjunto de componentes principais desejado.

3.3 Seleção de subconjuntos de variáveis

Embora o PCA seja comumente utilizado como etapa de parametrização e redução de dimensão de conjuntos de dados complexos, a utilização dos componentes principais em classificadores diretamente nem sempre é ideal.

A necessidade de preparação dos vetores de entrada para aplicação do PCA evidencia que a interpretação da influência dos componentes principais na classificação pode se tornar um processo trabalhoso. A identificação de um subconjunto de variáveis do conjunto original $\mathbf{x}$ permitiria verificar diretamente as intervenções no processo monitorado para correção de problemas.

Duas abordagens são apresentadas para atender a este fim. A primeira consiste na inspeção da matriz de transformação $\mathbf{A}$, que procura identificar os elementos de $\mathbf{x}$ que possuem maior influência para cada componente principal individualmente. A segunda abordagem, denominada *Principal Feature Analysis* (PFA), embora também inspecione a matriz $\mathbf{A}$, busca o melhor conjunto de variáveis que atendam um grupo de q componentes principais conjuntamente.

3.3.1 Inspeção da matriz de transformação

Um método intuitivo para identificação de variáveis é a análise das colunas da matriz de coeficientes $\mathbf{A}$. Os elementos da coluna j de $\mathbf{A}$ refletem os pesos que cada variável de $\mathbf{x}$ possui no j-ésimo componente principal. As p variáveis com os maiores pesos, em módulo, são escolhidas.

Ao passo que [16, 18] mostram que esta seleção não representa necessariamente o subconjunto com menor erro quadrático médio, o artigo [13] indica que esta técnica pode produzir bons resultados.

3.3.2 PFA - *Principal Feature Analysis*

O PFA, conforme descrito no artigo [14], seleciona p variáveis examinando, de forma simultânea, as relações de seus pesos na matriz $\mathbf{A}$ referentes aos q primeiros componentes principais, selecionados a partir de um valor escolhido de V_{retida}. Ainda de acordo com o artigo, a escolha de $p > q$, onde tipicamente $1 \leq p - q \leq 5$, pode ser necessária para garantir boa representatividade.

A sequência de procedimentos descrita a seguir define o PFA:

1. Aplicação do PCA em $\mathbf{x}$;
2. Seleção de q componentes principais que atendam ao valor de V_{retida} escolhido;
3. Obtenção da matriz $\mathbf{A}_q$ que contém as q primeiras colunas de $\mathbf{A}$; A i-ésima linha das n linhas da matriz $\mathbf{A}_q$ é denominada como $\mathbf{v}_i$;
4. Escolhem-se p grupos de vetores $\mathbf{v}$ através do algoritmo k-médias com a métrica de distância euclidiana;
5. Para cada grupo, escolhe-se a variável de $\mathbf{x}$ que corresponda ao vetor $\mathbf{v}$ mais próximo do centroide do grupo.

Esta técnica apresentou bons resultados em análises de imagens para detecção de características relevantes no reconhecimento de expressões faciais e para seleção de imagens com base em seu conteúdo [14]. Posteriormente, o PFA também se mostrou útil para análise de imagens de ressonância magnética [15].

3.4 Processamento e análise

O PCA é utilizado tanto nos conjuntos de dados apresentados na seção 2.4, onde são consideradas as amostras referentes a 24 horas de operação ininterrupta de 22 variáveis de processo, como também no conjunto de dados estendido, que contém todas as 86 variáveis de processo disponíveis. Conforme a teoria apresentada nas seções 3.2 e 3.3, estes dados devem ser pré-processados. Após este tratamento, são realizados dois estudos: o primeiro deles consiste na avaliação do número de componentes principais (CP) suficientes para representação dos casos de PCF e PSF; o segundo utiliza o PFA para selecionar as variáveis de entrada para os classificadores do Experimento 2 da seção 2.4.3.

3.4.1 Pré-processamento

Para aplicação do PCA no conjunto de dados de funcionamento dos turbogeradores, ressaltam-se as seguintes características observadas:

- A estatística dos sinais observados é desconhecida;
- Conforme Tabela 2.2, sabe-se que os sinais monitorados possuem medições em diferentes escalas.

De acordo com a seção 3.2, um conjunto de dados com estas propriedades deve ter sua estatística aproximada pelas equações (3.4) e (3.5). Adicionalmente, para impedir que as medições de diferentes unidades de medição e escalas influenciem negativamente, nos grupos de casos analisados, todos os sinais são processados para possuir média zero e normalizados pelos seus desvios-padrão.

Ao analisar a natureza das variáveis da Tabela 2.2, destaca-se que as *tags* KT-001/002, QT-001/002 e GT-001 representam contadores do número horas de operação, contadores da quantidade de partidas dos turbogeradores e o indicador do tipo de combustível utilizado, respectivamente. Esses sinais não possuem variação de estatística significativa para seu uso no PCA. A utilização da sequência temporal completa destas variáveis traz um excesso de informação desnecessário ao processo de treinamento de classificadores. Com isso, somente as 81 variáveis restantes são consideradas para aplicação do PFA.

3.4.2 Número de CPs para representação

Conforme indicado por [14], a quantidade de variáveis no domínio original suficiente para representação de um conjunto de sinais pode ser estimada a partir do seu número de componentes principais. O estudo apresentado nesta seção propõe indicar a quantidade de CPs mínima para representação dos conjuntos de dados analisados e, com isso, estimar o número mínimo de variáveis adequado para sua representação.

Como existem dois casos de paradas estudados, as PCFs e PSFs, avalia-se a variação do número componentes principais isoladamente e em conjunto. Com esta finalidade, estudou-se a variação dos resultados do PCA dentro dos seguintes subconjuntos:

- Casos isolados de PSF: para cada um dos 193 casos de PSF, a aproximação da estatística das variáveis foi realizada utilizando suas séries de 1440 amostras;
- Casos de PSF em conjunto: o cálculo do número de componentes principais é realizado utilizando todas as 1440 amostras dos 193 casos de PSF em conjunto, totalizando 277920 amostras por série temporal, para a aproximação da estatística das variáveis;
- Casos isolados de PCF: para cada um dos 33 casos de PCF, a aproximação da estatística das variáveis foi realizada utilizando suas séries de 1440 amostras;

- Casos de PCF em conjunto: o cálculo do número de componentes principais é realizado utilizando todas as 1440 amostras dos 33 casos de PCF em conjunto, totalizando 47520 amostras por série temporal, para a aproximação da estatística das variáveis;

- Todos os casos de PSF e PCF em conjunto: o cálculo do número de componentes principais é realizado utilizando todas as 1440 amostras de todos os 226 casos disponíveis, totalizando 325440 amostras, para a aproximação da estatística das variáveis.

• Conjunto de 22 variáveis

Adotando como premissa a retenção de 90% de variabilidade retida, aplicou-se o PCA nas variáveis da Tabela 2.4. Os resultados são apresentados conforme o agrupamento de casos proposto.

◇ Casos isolados de PSF

Neste conjunto de dados, o número de componentes principais (CP) variou entre 1 e 3 CPs. A Tabela 3.1 apresenta a quantidade de casos, por turbogerador, que é representada pelo respectivo número de CPs. Verifica-se que, utilizando 3 CPs, consegue-se representar isoladamente cada caso de PSF retendo no mínimo 90% de sua energia. Observa-se também que os casos analisados de PSF do turbogerador TGA apresentam concentração de energia maior, podendo ser representados por, no mínimo, 2 CPs.

Tabela 3.1: PSF, 22 variáveis: componentes principais por turbogeradores.

	TGA	**TGB**	**TGC**	**TGD**
PSFs com CP=1	31	36	43	28
PSFs com CP=2	18	9	4	15
PSFs com CP=3	0	2	1	5

◇ Casos isolados de PCF

Neste conjunto de dados, o número de componentes principais também variou entre 1 e 3 CPs. A Tabela 3.2 apresenta a quantidade de casos, por turbogerador, que é representada pelo respectivo número de CPs. Verifica-se que, novamente, utilizando 3 CPs, consegue-se representar isoladamente cada caso de PCF retendo no mínimo 90% de sua energia. Observa-se também que os casos analisados de PCF dos turbogeradores TGB e TGC apresentam concentração de energia maior, podendo ser representados por, no mínimo, 2 CPs.

Tabela 3.2: PCF, 22 variáveis: componentes principais por turbogeradores.

	TGA	**TGB**	**TGC**	**TGD**
PCFs com CP=1	4	5	8	4
PCFs com CP=2	1	4	2	3
PCFs com CP=3	1	0	0	1

◇ Agrupamento de PSFs, PCFs e conjunto de PSFs e PCFs

O número de componentes principais para os dados agrupados das PSFs, PCFs e do conjunto de PSFs e PCFs é apresentado na Tabela 3.3. Verifica-se que utilizando 5 CPs, consegue-se representar os conjuntos de PCFs e PSFs retendo no mínimo 90% de sua energia. O aumento do número de CPs é decorrente da maior complexidade para representação, que se deve ao aumento do conjunto para análise estatística em comparação com a análise de casos de parada isolados.

Tabela 3.3: Grupos de PSFs e PCFs, 22 variáveis: componentes principais.

	Componentes Principais
Grupo PSF	4
Grupo PCF	5
Grupo PSF+PCF	5

• Conjunto de 81 variáveis

Adotando como premissa a retenção de 90% de variabilidade retida, aplicou-se o PCA nas variáveis da Tabela 2.2. Os resultados são apresentados conforme o agrupamento de casos proposto.

◇ Casos isolados de PSF

Neste conjunto de dados, o número de componentes principais variou entre 5 e 25 CPs. A Tabela 3.4 apresenta a quantidade de casos, por turbogerador, que é representada pelo respectivo número de CPs. Verifica-se que o aumento do número de variáveis trouxe a necessidade de uma quantidade maior de CPs para representação dos casos de PSF mantendo 90% da energia retida. Como a energia está mais distribuída, considerando o caso de maior distribuição de energia entre as CPs do turbogerador TGC, são necessárias 25 CPs para representar isoladamente cada caso de PSF.

◇ Casos isolados de PCF

Neste conjunto de dados, o número de componentes principais variou entre 7 e 25 CPs. A Tabela 3.5 apresenta a quantidade de casos, por turbogerador, que é representada pelo respectivo número de CPs. Verifica-se que o aumento do número de

Tabela 3.4: PSF, 81 variáveis: componentes principais por turbogeradores.

	TGA	**TGB**	**TGC**	**TGD**
PSFs com CP=5	0	1	0	1
PSFs com CP=6	2	0	0	0
PSFs com CP=7	2	1	2	4
PSFs com CP=8	1	2	0	3
PSFs com CP=9	5	1	6	2
PSFs com CP=10	3	4	6	5
PSFs com CP=11	4	4	9	2
PSFs com CP=12	4	8	8	5
PSFs com CP=13	7	6	4	5
PSFs com CP=14	2	2	3	4
PSFs com CP=15	5	6	1	3
PSFs com CP=16	2	2	4	3
PSFs com CP=17	2	2	0	1
PSFs com CP=18	1	4	1	2
PSFs com CP=19	7	3	0	2
PSFs com CP=20	0	0	1	2
PSFs com CP=21	1	0	1	2
PSFs com CP=22	1	1	0	1
PSFs com CP=23	0	0	1	0
PSFs com CP=25	0	0	1	0

variáveis trouxe a necessidade de uma quantidade maior de CPs para representação dos casos de PCF mantendo 90% da energia retida. Como a energia está mais distribuída, considerando o caso de maior distribuição de energia entre as CPs do turbogerador TGA, são necessárias 25 CPs para representar isoladamente cada caso de PCF.

Tabela 3.5: PCF, 81 variáveis: componentes principais por turbogeradores.

	TGA	**TGB**	**TGC**	**TGD**
PCFs com CP=7	0	0	1	0
PCFs com CP=9	0	3	0	0
PCFs com CP=10	1	0	0	0
PCFs com CP=11	0	0	3	0
PCFs com CP=12	0	1	1	1
PCFs com CP=13	2	1	0	1
PCFs com CP=14	1	0	1	0
PCFs com CP=15	0	1	1	1
PCFs com CP=16	0	2	1	1
PCFs com CP=17	0	0	1	2
PCFs com CP=18	1	0	1	1
PCFs com CP=19	0	1	0	1
PCFs com CP=25	1	0	0	0

◇ Agrupamento de PSFs, PCFs e conjunto de PSFs e PCFs

O número de componentes principais para os dados agrupados das PSFs, PCFs e do conjunto de PSFs e PCFs é apresentado na Tabela 3.6. Verifica-se que, utilizando

16 CPs, consegue-se representar os conjuntos de PCFs e PSFs retendo no mínimo 90% de sua energia. Neste caso, o aumento do conjunto para análise estatística em comparação com a análise de casos de parada isolados permitiu que o número de CPs seja menor que a quantidade utilizada para a representação individual. Ainda assim, o número de variáveis maior resultou num aumento significativo do número de CPs necessário para representação do conjunto de PCFs e PSFs em comparação com o conjunto de 22 variáveis.

Tabela 3.6: Grupos de PSFs e PCFs, 81 variáveis: componentes principais.

	Componentes Principais
Grupo PSF	16
Grupo PCF	14
Grupo PSF+PCF	16

3.4.3 Seleção de variáveis de turbogeradores

Os resultados obtidos na seção 3.4.2 fornecem uma estimativa inicial para o número de variáveis que devem ser utilizadas para treinamento de novos classificadores. O objetivo desta seção é, através da realização do Experimento 2 da seção 2.4.3, determinar novos conjuntos de variáveis para comparação com o conjunto de 22 variáveis utilizado em [3].

Dentre os métodos apresentados na seção 3.3, o PFA foi escolhido em detrimento da inspeção da matriz de transformação.

A partir da verificação das matrizes de transformação geradas, notou-se que os pesos relativos das variáveis possuem valores muito próximos. Logo, a determinação de uma métrica para definir o número de variáveis que devem ser mantidas mostrou-se ineficiente. Soma-se ainda a dificuldade para relacionar o número de variáveis com a quantidade de componentes principais identificadas e suas importâncias relativas.

Por outro lado, o PFA apresenta uma metodologia que determina o número de variáveis considerando as relações do conjunto de componentes principais mantidas. Da mesma forma que o artigo [14], foi utilizado o algoritmo k-médias minimizando a distância euclidiana. No entanto, não foi discutido como minimizar a possibilidade de utilizar grupos que correspondam a mínimos locais. A estratégia proposta neste trabalho consiste em escolher o resultado que possua o menor somatório das distâncias dos vetores $\mathbf{v}_i$ mais próximos aos centroides dos grupos determinados pela execução de 1000 repetições, com inicializações distintas, do algoritmo k-médias.

- **Aplicação do PFA no Experimento 2**

Diferentemente do estudo da seção 3.4.2, que utilizou todos os casos disponíveis de PSFs e PCFs, para a realização do Experimento 2, a aplicação do PFA deve contar

somente com os dados disponíveis durante a fase de treinamento dos classificadores. Assim, o PFA foi utilizado em cada uma das 20 combinações de dados de treinamento disponíveis no Experimento 2.

As estatísticas das variáveis foram modeladas através dos grupos de PSFs, de PCFs e do agrupamento total de PCFs e PSFs, partindo do conjuntos de 22 e 81 variáveis. O número de variáveis selecionadas variou entre a quantidade de CPs identificada, para dada combinação, e a mesma acrescida de até 5 variáveis adicionais, conforme proposto em [14]. Geraram-se 36 blocos de 20 conjuntos de variáveis para cada combinação de modelagem estatística, número de variáveis selecionadas e conjuntos de variáveis originais propostos.

• Métrica de avaliação das variáveis selecionadas

Para cada conjunto de variáveis de cada combinação de grupos, são treinados classificadores conforme seção 2.4.3.

A métrica utilizada para avaliação de desempenho dos classificadores são seus F1-*Score* associados. Conforme descrito em [19], o F1-*Score* é a média harmônica entre a precisão e a sensibilidade de um classificador.

Para a representação do cálculo do F1-*Score*, primeiramente, definimos os seguintes termos:

- $\mathrm{PCF}_{\mathrm{corr}}$: PCFs identificadas corretamente (verdadeiros positivos);
- $\mathrm{PCF}_{\mathrm{incorr}}$: PCFs identificadas incorretamente (falsos negativos);
- $\mathrm{PSF}_{\mathrm{corr}}$: PSFs identificadas corretamente (verdadeiros negativos);
- $\mathrm{PSF}_{\mathrm{incorr}}$: PSFs identificadas incorretamente (falsos positivos).

Com isso, o F1-*Score* é calculado por

$$F1\text{-}Score = 2 \times \frac{\text{precisão} \times \text{sensibilidade}}{\text{precisão} + \text{sensibilidade}}, \tag{3.9}$$

onde

$$\text{precisão} = \frac{\mathrm{PCF}_{\mathrm{corr}}}{\mathrm{PCF}_{\mathrm{corr}} + \mathrm{PSF}_{\mathrm{incorr}}} \tag{3.10}$$

e

$$\text{sensibilidade} = \frac{\mathrm{PCF}_{\mathrm{corr}}}{\mathrm{PCF}_{\mathrm{corr}} + \mathrm{PCF}_{\mathrm{incorr}}}. \tag{3.11}$$

Ao analisar a equação (3.9), percebe-se que os valores da F1-*Score* variam entre $[0, 1]$. Enquanto o valor 1 representa a classificação sem erros, o valor nulo equivale à ausência de identificação de PCFs. O valor de referência para comparação de resultados é igual a 0,7212, que é o F1-*Score* médio obtido no Experimento 2.

Esta métrica, utilizada comumente na literatura [20], foi escolhida para simplificar a análise dos acertos dos classificadores, que foi apresentada no Capítulo 2 separadamente por taxas de acerto médio de PCFs e de PSFs.

- **Conjunto de 22 variáveis**

O Experimento 2 foi realizado com todos os 18 blocos de conjuntos de variáveis obtidos pela combinação das 3 modelagens estatísticas propostas com os conjuntos de variáveis compostos pelo número de componentes principais com 0 a 5 variáveis adicionais, partindo do conjunto de 22 variáveis original. A comparação de desempenho dos classificadores é realizada através do valor da média do F1-*Score* obtido nas 5 repetições das 20 combinações dos 6 grupos de casos do Experimento 2.

Na Figura 3.1, observa-se que 2 blocos de conjuntos de variáveis alcançaram resultados melhores que o Experimento 2 original. O bloco de melhor desempenho é composto pelos conjuntos de variáveis, com 5 variáveis adicionais, selecionados através da modelagem estatística a partir de casos de PCFs. A quantidade total de variáveis destes conjuntos variou entre 9 e 11. O valor do F1-*Score* médio deste conjunto é 0,7231. Este valor corresponde a uma taxa de acerto médio de 85,45% para as PSFs e de 74,66% para as PCFs.

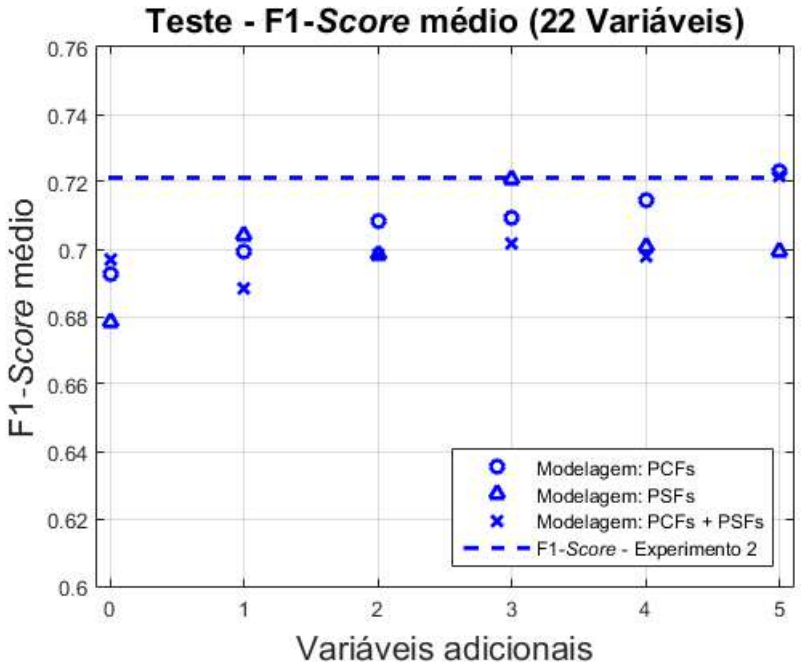


Figura 3.1: Com a seleção a partir do conjunto de 22 variáveis, observa-se que 2 blocos de conjuntos de variáveis alcançaram resultados melhores que o Experimento 2. O F1-*Score* médio do conjunto de melhor desempenho é 0,7231.

Verificou-se que as variáveis selecionadas variaram bastante para cada combinação de grupos. A quantidade de combinações em que cada uma das variáveis

foi selecionada é apresentada na Figura 3.2. Observa-se que somente uma variável, TI-008 (sensor de temperatura na exaustão), não foi selecionada em nenhuma das combinações. A Tabela 3.7 lista as variáveis presentes em mais de 50% das combinações de grupos.

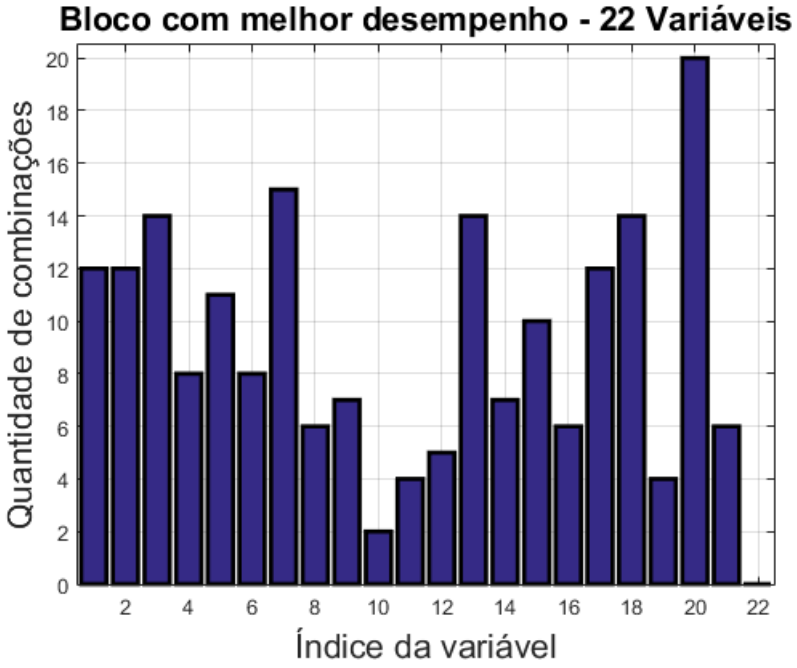


Figura 3.2: No bloco de conjuntos de variáveis, selecionadas das 22 disponíveis, com melhor desempenho médio, observa-se que 10 variáveis, listadas na Tabela 3.7, estão presentes em 10 ou mais combinações de grupos. Somente o TI-008 (sensor de temperatura na exaustão) não foi utilizado em nenhuma das combinações.

Tabela 3.7: PFA, 22 variáveis: variáveis presentes em mais de 50% das combinações.

Índice	Identificação	Descrição	Combinações
1	TI-003-01	Termopar 01 do perfil da exaustão	12
2	TI-003-02	Termopar 02 do perfil da exaustão	12
3	TI-003-03	Termopar 03 do perfil da exaustão	14
5	TI-003-05	Termopar 05 do perfil da exaustão	11
7	TI-003-07	Termopar 07 do perfil da exaustão	15
13	TI-003-13	Termopar 13 do perfil da exaustão	14
17	TI-003-17	Termopar 17 do perfil da exaustão	12
18	FIT-001	Vazão de Gás Combustível	14
20	PDT-001	Dif. de pressão na entrada do TG	20

A Figura 3.5(a) mostra a quantidade de erros de predição por caso de PCF, separados por seu tipo, quando presentes no conjunto de testes, referentes ao bloco de conjunto de variáveis com melhor desempenho médio. Considerando que cada PCF está presente no conjunto de testes em 10 combinações de grupos e que cada experimento é repetido 5 vezes, o número máximo de erros de predição possível é igual a 50. Observa-se que um caso de GDF foi identificado incorretamente em todos

classificadores treinados. Além disso, verifica-se que um total de 2 GDFs, 2 GUFs e 2 OIFs são identificadas incorretamente em mais de 50% das oportunidades em que foram testadas.

- **Conjunto de 81 variáveis**

O Experimento 2 foi realizado com todos os 18 blocos de conjuntos de variáveis obtidos pela combinação das 3 modelagens estatísticas propostas com os conjuntos de variáveis compostos pelo número de componentes principais com 0 a 5 variáveis adicionais, partindo do conjunto de 81 variáveis. A comparação de desempenho dos classificadores é realizada através do valor da média do F1-*Score* obtido nas 5 execuções das 20 combinações dos 6 grupos de casos do Experimento 2.

Na Figura 3.3, observa-se que 8 blocos de conjuntos de variáveis alcançaram resultados melhores que o Experimento 2 original e que um bloco igualou este resultado. O bloco de melhor desempenho é composto pelos conjuntos de variáveis, sem variáveis adicionais, selecionados através da modelagem estatística a partir de casos de PSFs. A quantidade total de variáveis destes conjuntos variou entre 11 e 13. O valor do F1-*Score* médio deste conjunto é 0,7425. Este valor corresponde a uma taxa de acerto médio de 84,43% para as PSFs e de 77,96% para as PCFs.

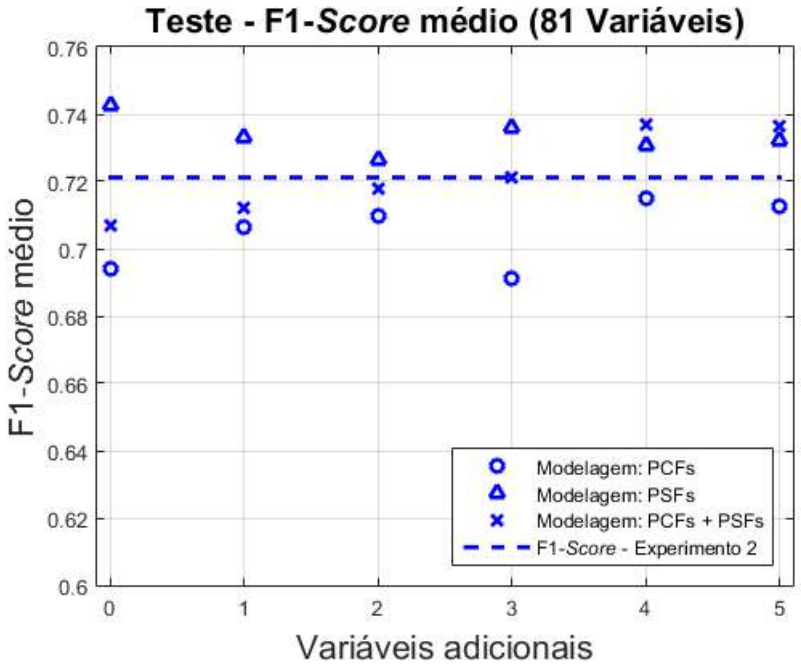


Figura 3.3: Com a seleção a partir do conjunto de 81 variáveis, observa-se que 8 blocos de conjuntos de variáveis alcançaram resultados melhores que o Experimento 2 original e que um bloco igualou este resultado. O F1-*Score* médio do conjunto de melhor desempenho é 0,7425.

Verificou-se que, novamente, as variáveis selecionadas variaram bastante para cada combinação de grupos. A quantidade de combinações em que cada uma das

variáveis foi selecionada é apresentada na Figura 3.4. A Tabela 3.8 lista as 2 variáveis presentes em mais de 50% das combinações de grupos. Neste bloco, 28 variáveis, listadas na Tabela 3.9, não foram selecionadas em nenhuma das combinações.

Tabela 3.8: PFA, 81 variáveis: variáveis presentes em mais de 50% das combinações.

Índice	**Identificação**	**Descrição**	**Combinações**
69	TI-007	Gerador - temperatura enrolamento L1-31	11
70	TI-008	Exaustão GG	16

Tabela 3.9: PFA, 81 variáveis: variáveis não selecionadas nas combinações.

Índice	**Identificação**	**Descrição**
11	VE-003	Vibração na turbina do GG
13	VE-005	Vib. PT disc. end. Y
15	VE-007	Vib. PT acopl. Y
20	TI-003-01	Termopar 01 do perfil da exaustão
21	TI-003-01	Termopar 02 do perfil da exaustão
22	TI-003-01	Termopar 03 do perfil da exaustão
23	TI-003-01	Termopar 04 do perfil da exaustão
24	TI-003-01	Termopar 05 do perfil da exaustão
25	TI-003-01	Termopar 06 do perfil da exaustão
27	TI-003-01	Termopar 08 do perfil da exaustão
33	TI-003-01	Termopar 14 do perfil da exaustão
34	TI-003-01	Termopar 15 do perfil da exaustão
37	TI-004	Gerador temp. enrolamento L2-3
38	TI-005	Gerador temp. enrolamento L2-2
39	TI-006	Gerador temp. enrolamento L2-1
45	PT-001	Pressão P1
52	ST-004	Rotação PT
53	TE-001	Exaustão GG
54	IT-001	Corrente de excitação do campo
59	JT-002	*Active Power*
68	IT-002	Corrente
73	TI-011	PT Temperatura mancal DE
76	TI-014	Gearbox LSS Temperatura mancal NDE
77	TI-015	Gearbox HSS Temperatura mancal NDE
79	TI-017	Gerador ar de resfriamento (quente) DE
80	TI-018	Gerador ar de resfriamento (quente) NDE
81	TI-019	Gerador ar de resfriamento (frio) NDE
83	TI-021	Gerador temp. enrolamento L1-1

A Figura 3.5(b) mostra a quantidade de erros de predição por caso de PCF, separados por seu tipo, quando presentes no conjunto de testes, referentes ao bloco de conjunto de variáveis com melhor desempenho médio. Observa-se que um caso de GUF foi identificado incorretamente em todos classificadores treinados. Além disso, verifica-se que um total de 2 GDFs, 3 GUFs e 1 OIF são identificadas incorretamente em mais de 50% dos oportunidades em que foram testadas. Em comparação com a Figura 3.5(a), nota-se que, tanto para o conjunto de GDFs como para o conjunto de OIFs, o número de erros médio foi atenuado e que os erros se distribuíram mais entre

os casos. Essa tendência não foi observada para o conjunto de GUFs, que, embora tenha apresentado um desempenho médio melhor, manteve 4 casos com alta taxa de erros de predição.

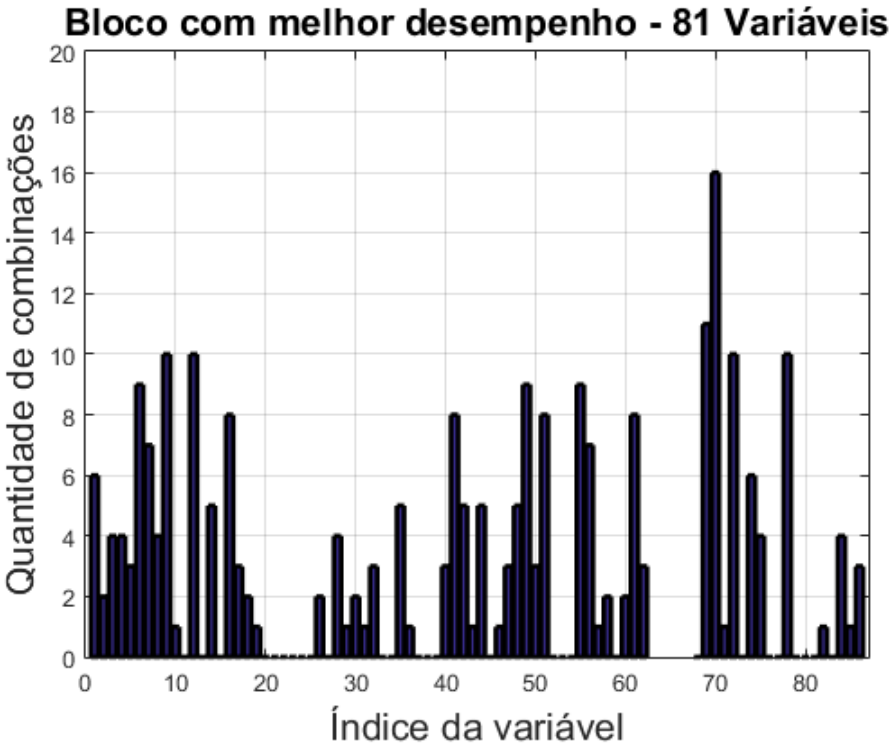


Figura 3.4: No bloco de conjuntos de variáveis, selecionadas das 81 disponíveis, com melhor desempenho médio, observa-se que 2 variáveis, listadas na Tabela 3.8, estão presentes em mais de 10 combinações de grupos. Há, também, 28 variáveis, listadas na Tabela 3.9, não selecionadas em nenhuma das combinações.

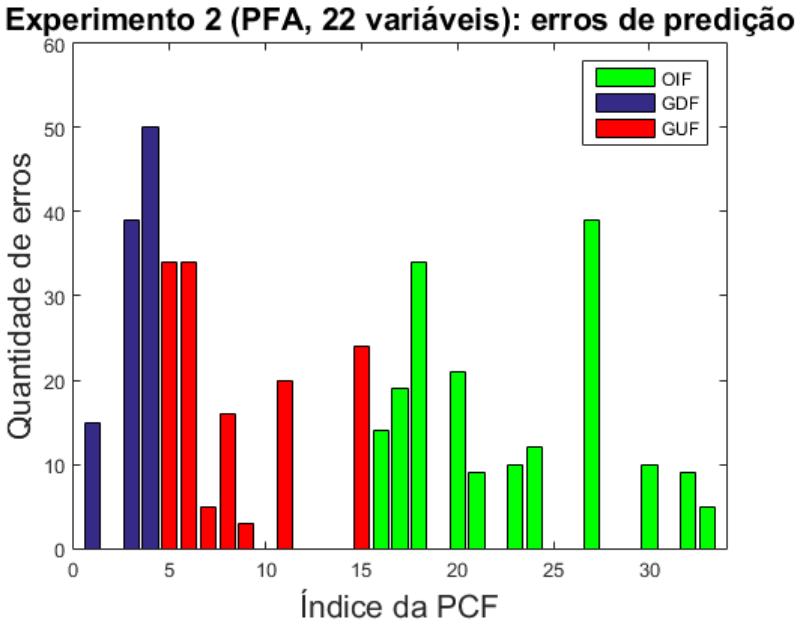
3.5 Conclusão

Este capítulo aprofundou o estudo do banco de dados utilizado em [3]. Utilizando a técnica de PCA, conforme [16, 17], analisou-se a quantidade de CPs necessárias para representar os conjuntos de PSFs e PCFs utilizados no Experimento 2, conforme estudo apresentado na seção 3.4.2. O número de variáveis analisadas foi estendido para um conjunto total de 81 variáveis.

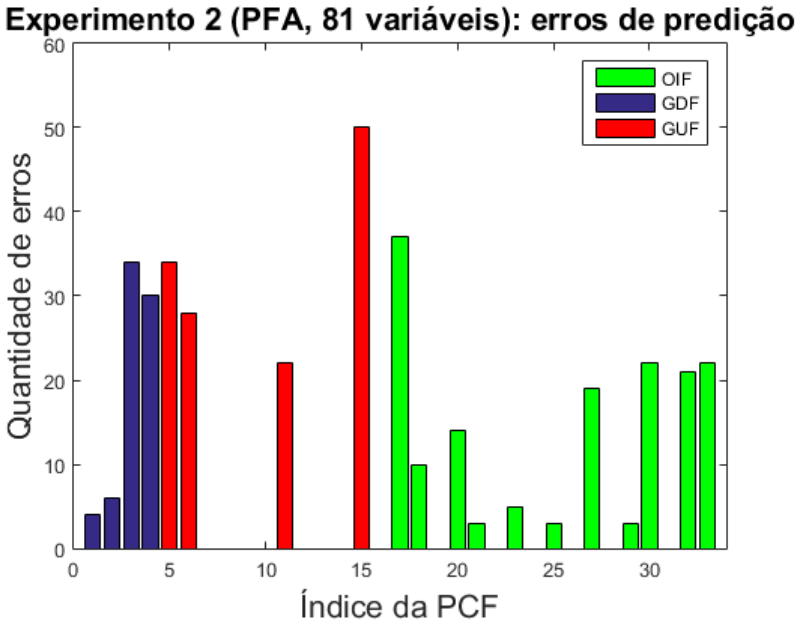
Procurando estabelecer uma relação entre o número de CPs que representam, com 90% da variabilidade retida, o conjunto de dados original e a quantidade de variáveis de processo que seriam necessárias para mesma representatividade, o estudo da seção 3.4.3 foi realizado. Utilizou-se a técnica PFA [14], que possui sua base na teoria do PCA e no uso do algoritmo de k-médias. Com as variáveis selecionadas, modeladas pelas estatísticas dos conjuntos de PCFs e PFSs dos dados de treinamento de cada uma das 20 combinações de grupos disponíveis, foram obtidos resultados

equivalentes ao Experimento 2 reduzindo o conjunto das 22 variáveis de entrada, utilizadas em [3], para subconjuntos de 9 a 11 variáveis. Além disso, reduzindo o conjunto estendido de 81 variáveis para subconjuntos entre 11 e 13 variáveis, obteve-se resultado superior ao Experimento 2 executado no Capítulo 2. Desta forma, desenvolveu-se uma técnica de seleção de variáveis semisupervisionada efetiva para o problema apresentado.

Observando-se os resultados de identificação de PCFs, na etapa de teste dos classificadores, apresentados nas Figuras 3.5(a) e 3.5(b), percebe-se que ainda há casos que não foram identificados adequadamente com este conjunto de variáveis selecionado. Considerando que o PCA é capaz de identificar somente as relações lineares entre as variáveis de um conjunto, deve-se também avaliar se há relações não lineares que permitam uma melhor identificação dos casos de PSF e PCF. No Capítulo 4, técnicas para mapeamento destas relações serão estudadas.



(a)



(b)

Figura 3.5: Erros de predição em casos de PCF: (a) observa-se que um caso de GDF foi identificado incorretamente em todos classificadores treinados. Além disso, verifica-se que um total de 2 GDFs, 2 GUFs e 2 OIFs são identificadas incorretamente em mais de 50% das oportunidades em que foram testadas; (b) observa-se que um caso de GUF foi identificado incorretamente em todos classificadores treinados. Verifica-se que total de 2 GDFs, 3 GUFs e 1 OIF são identificadas incorretamente em mais de 50% dos oportunidades em que foram testadas. Em comparação com (a), nota-se que, tanto para o conjunto de GDFs como para o conjunto de OIFs, o número de erros médio foi atenuado e que os erros se distribuíram mais entre os casos. Essa tendência não foi observada para o conjunto de GUFs, que, embora tenha melhorado o desempenho médio, manteve 4 casos com alta taxa de erros de predição.

Capítulo 4

Relações Não Lineares em Domínios Transformados

4.1 Introdução

No capítulo 3, ao utilizar o PCA, obteve-se uma metodologia capaz de realizar a seleção de variáveis de forma semisupervisionada. No entanto, o PCA é capaz de detectar de forma eficiente somente relações lineares entre as variáveis que compõem um conjunto de dados.

Considerando a complexidade do funcionamento de um turbogerador, espera-se que haja também relações não lineares dentro de suas grandezas de processo monitoradas. Logo, a capacidade de identificação destas relações pode permitir a obtenção de novas informações que sejam relevantes para a identificação dos casos de PCF e PSF estudados.

Uma possível solução é trabalhar com domínios transformados, compostos por combinações não lineares das variáveis originais. Conforme discutido em [21, 22], realizar explicitamente a transformação de domínios pode representar um aumento demasiado de custo computacional. Para contornar esta limitação, conforme apresentado em [21, 23–26], utilizam-se as funções de *kernel*. Com elas, é possível calcular algumas operações, dentre elas o PCA, em domínios transformados sem realizar explicitamente a mudança de domínio.

Na seção 4.2, são apresentadas a aplicação do PCA em domínios transformados e sua implementação alternativa utilizando funções de *kernel*, o *Kernel Principal Component Analysis* (KPCA). Através dela, é possível obter CPs em domínios transformados não linearmente de acordo com o tipo de *kernel* escolhido. A utilização do KPCA nos dados dos turbogeradores e os resultados obtidos são apresentados na seção 4.3. Por fim, a seção 4.4 sintetiza os estudos realizados e indica as principais conclusões das análises propostas.

4.2 PCA em domínios transformados

Nesta seção, o PCA, conforme apresentado no capítulo 3, será aplicado em domínios transformados.

Dados M conjuntos de vetores $\mathbf{x}$, conforme seção 3.2, define-se uma transformação Φ [21], através de

$$\Phi : \begin{cases} \mathbb{R}^N \to F \\ \mathbf{x} \to \mathbf{X}. \end{cases} \tag{4.1}$$

O espaço F é denominado espaço de *features*. Ele possui dimensionalidade arbitrariamente grande, podendo ser, inclusive, infinita.

Para os cálculos a seguir, assume-se que os dados no domínio transformado possuem média igual a zero. Para aplicação do PCA em F, é utilizada a matriz de covariância amostral $\mathbf{S_X}$, que é calculada pela expressão

$$\mathbf{S_X} = \frac{1}{M} \sum_{i=1}^{M} \Phi(\mathbf{x}_i)\Phi^T(\mathbf{x}_i). \tag{4.2}$$

Para o levantamento das CPs no domínio transformado, deve-se solucionar as equações para encontrar os autovalores λ e autovetores $\mathbf{w}$ de $\mathbf{S_X}$ que satisfaçam

$$\lambda \mathbf{w} = \mathbf{S_X} \mathbf{w}. \tag{4.3}$$

Pode-se verificar que $\mathbf{S_X w} = \frac{1}{M} \sum_{i=1}^{M} (\Phi(\mathbf{x}_i) \cdot \mathbf{w}) \Phi(\mathbf{x}_i)$, onde $(\Phi(\mathbf{x}_i) \cdot \mathbf{w})$ representa o produto interno entre $\Phi(\mathbf{x}_i)$ e $\mathbf{w}$. Com isso, a equação (4.3) pode ser também representada por

$$\lambda(\Phi(\mathbf{x}_k) \cdot \mathbf{w}) = (\Phi(\mathbf{x}_k) \cdot \mathbf{S_X w}), \text{ para } k = 1, ..., M, \tag{4.4}$$

e sabe-se que existem coeficientes α_i com i variando de 1 a M tal que

$$\mathbf{w} = \sum_{i=1}^{M} \alpha_i \Phi(\mathbf{x}_i). \tag{4.5}$$

Combinando as equações (4.4) e (4.5), obtém-se a igualdade

$$\lambda \sum_{i=1}^{M} \alpha_i (\Phi(\mathbf{x}_k) \cdot \Phi(\mathbf{x}_i)) = \frac{1}{M} \sum_{i=1}^{M} \alpha_i (\Phi(\mathbf{x}_k) \cdot \sum_{j=1}^{M} \Phi(\mathbf{x}_j))(\Phi(\mathbf{x}_j) \cdot \Phi(\mathbf{x}_i)), \tag{4.6}$$

para $k = 1, ..., M$.

Definindo-se, também, uma matriz $\mathbf{K}$ quadrada de dimensão M tal que

$$\mathbf{K}(i,j) = k(\mathbf{x}_i, \mathbf{x}_j) = (\Phi(\mathbf{x}_i) \cdot \Phi(\mathbf{x}_j)), \tag{4.7}$$

a equação (4.6) é condensada pela expressão

$$M\lambda\mathbf{K}\boldsymbol{\alpha} = \mathbf{K}^2\boldsymbol{\alpha}, \tag{4.8}$$

onde $\boldsymbol{\alpha}$ é o vetor dos M coeficientes α_i. Assumindo que $\mathbf{K}$ é inversível, a equação (4.8) simplificada se torna

$$M\lambda\boldsymbol{\alpha} = \mathbf{K}\boldsymbol{\alpha}. \tag{4.9}$$

A matriz $\mathbf{K}$, similarmente à $\mathbf{R}_x$ da seção 3.2, é uma matriz simétrica. Seus M autovalores são as soluções $M\lambda$ da equação (4.8). Os autovetores de $\mathbf{K}$, $\boldsymbol{\alpha}^1, ..., \boldsymbol{\alpha}^M$, a diagonalizam.

Para aplicação do PCA no domínio transformado, os autovetores $\mathbf{w}^k$, com k variando de 1 a M devem possuir módulo unitário, ou seja, $\|\mathbf{w}^k\|^2 = 1$. Utilizando a equação (4.5) nesta expressão, tem-se

$$1 = \sum_{i,j=1}^{M} \alpha_i^k \alpha_j^k K_{ij} = (\boldsymbol{\alpha}^k \cdot \mathbf{K}\boldsymbol{\alpha}^k) = \lambda_k(\boldsymbol{\alpha}^k \cdot \boldsymbol{\alpha}^k), \tag{4.10}$$

impondo, desta forma, uma condição de normalização para os vetores $\boldsymbol{\alpha}^k$. Esta normalização deve ser realizada pela multiplicação de $\boldsymbol{\alpha}^k$ por $\frac{1}{\sqrt{\lambda_k}}$.

Por fim, a extração das componentes principais é realizada através da projeção de novos dados nos autovetores $\mathbf{w}^k$. Desta forma, para um novo vetor de teste $\mathbf{x}$, a sua projeção é calculada por

$$(\mathbf{w}^k \cdot \Phi(\mathbf{x})) = \sum_{i=1}^{M} \alpha_i^k (\Phi(\mathbf{x}_i) \cdot \Phi(\mathbf{x})) = \sum_{i=1}^{M} \alpha_i^k k(\mathbf{x}_i, \mathbf{x}). \tag{4.11}$$

4.2.1 Utilização de funções *kernel*

Considerando que a dimensão de F pode ser arbitrariamente grande, ou até infinita, o custo computacional envolvido para realizar a transformação pode ser muito alto ou mesmo impossível. Uma alternativa, proposta inicialmente por [23], consiste em determinar *a priori* a matriz $\mathbf{K}$ sem se preocupar propriamente com a função Φ de mapeamento correspondente.

Conforme discutido em [24–27], se a função $k(\mathbf{x}, \mathbf{y})$ for um operador simétrico positivo definido, atendendo ao Teorema de Mercer, é garantido que há um mapeamento para produtos internos no domínio F. Esta condição se resume nos autova-

lores λ_k de $\mathbf{K}$ serem positivos [21].

Operadores indefinidos também podem ser utilizados para extração de características não lineares [21]. No entanto, como pode haver autovalores negativos, a normalização dos autovetores é modificada, sendo realizada pela multiplicação por $\frac{1}{\sqrt{|\lambda_k|}}$.

O desenvolvimento matemático apresentado, neste capítulo, para o cálculo do PCA destaca sua representação via produtos internos justamente para evidenciar que a sua execução em domínios transformados pode se beneficiar da utilização de funções de *kernel* sem a necessidade de realizar a transformação de domínios, também chamado de *kernel trick*.

Há uma grande variedade de *kernels* apresentados em [21, 25, 28]. Nesta dissertação dois tipos de *kernel* são utilizados:

- *Kernel* polinomial de ordem d: $k(\mathbf{x}, \mathbf{y}) = (\mathbf{x}^T\mathbf{y})^d$
- *Kernel* gaussiano, com desvio-padrão σ: $k(\mathbf{x}, \mathbf{y}) = e^{-\left(\frac{\|\mathbf{x}-\mathbf{y}\|}{2\sigma^2}\right)^2}$

4.2.2 Média zero no domínio transformado

No domínio das variáveis de entrada, garantir que os sinais possuam média zero é uma tarefa trivial. Contudo, no domínio transformado, as novas representações destes sinais podem possuir média diferente de zero. Sem realizar a transformação de domínios, o cálculo da média não pode ser realizado diretamente.

Com o uso de funções de *kernel*, a manutenção da média igual a zero é possível através do ajuste da matriz $\mathbf{K}$ original [21]. Assim, a matriz modificada $\tilde{\mathbf{K}}$, que garante média zero do conjunto de vetores $\Phi(\mathbf{x}_i)$, é obtida através da equação

$$\tilde{\mathbf{K}} = \mathbf{K} - \mathbf{1}_M\mathbf{K} - \mathbf{K}\mathbf{1}_M + \mathbf{1}_M\mathbf{K}\mathbf{1}_M, \tag{4.12}$$

onde $\mathbf{1}_M$ é uma matriz quadrada de tamanho M com seus valores iguais a $\frac{1}{M}$.

Da mesma forma, as projeções dos conjuntos de dados de testes também devem ser corrigidas. Supondo um conjunto de teste de L elementos, define-se $\mathbf{K}_{\text{teste}}$ como a matriz de dimensão $L \times M$ que contém o cálculo das funções de *kernel* entre os conjuntos de teste e de treino. A matriz com seus dados centralizados em torno da média dos dados de treino, $\tilde{\mathbf{K}}_{\text{teste}}$, é calculada através da expressão

$$\tilde{\mathbf{K}}_{\text{teste}} = \mathbf{K}_{\text{teste}} - \mathbf{1}'_M\mathbf{K} - \mathbf{K}_{\text{teste}}\mathbf{1}_M + \mathbf{1}'_M\mathbf{K}\mathbf{1}_M, \tag{4.13}$$

onde $\mathbf{1}'_M$ é uma matriz $L \times M$ com seus valores iguais a $\frac{1}{M}$.

4.2.3 KPCA - *Kernel* PCA

Considerando a teoria apresentada nas seções anteriores, o algoritmo do KPCA pode ser resumido nos seguintes passos:

1. Escolha da função de *kernel* $k(\mathbf{x}, \mathbf{y})$ e o cálculo da matriz $\mathbf{K}$ para os M elementos do conjunto de treino;
2. Cálculo da matriz $\mathbf{K}$ e $\mathbf{K}_{\text{teste}}$ através da função de *kernel*;
3. Retirada da média no domínio transformado calculando $\tilde{\mathbf{K}}$ e $\tilde{\mathbf{K}}_{\text{teste}}$ conforme equações (4.12) e (4.13);
4. Obter os autovetores $\tilde{\boldsymbol{\alpha}}_k$ de $\tilde{\mathbf{K}}$ e realizar a normalização deles multiplicando-os por $\dfrac{1}{\sqrt{\tilde{\lambda}_k}}$;
5. Realizar as projeções de $\tilde{\mathbf{K}}_{\text{teste}}$ nos vetores $\tilde{\boldsymbol{\alpha}}_k$ e obter as componentes principais no domínio transformado, utilizando a equação (4.11).

O KPCA mantém todas as propriedades do PCA no domínio transformado, desde que a escolha da função de *kernel* atenda aos requisitos discutidos na seção 4.2.1.

Uma das principais diferenças entre o PCA e o KPCA está no número de características que podem ser obtidas. No PCA, independente do número de elementos do conjunto utilizado no treinamento, o número máximo de características extraídas é igual à quantidade de variáveis constantes em $\mathbf{x}$. Já no KPCA, este valor passa a ser igual ao número de elementos M do conjunto de treino. Além disso, utilizando funções de *kernel* diferentes, esse número pode ser maior que M.

Em termos de processamento, o KPCA exige que o conjunto de treinamento seja utilizado para o cálculo das projeções de cada elemento de teste. No PCA, uma vez calculada a matriz de transformação, os dados de treinamento não são mais utilizados.

4.3 Processamento e análise

O KPCA é aplicado no conjunto de dados estendido, utilizando as 81 variáveis de processo selecionadas na seção 3.4.1. O condicionamento dos sinais de entrada para as aplicações do KPCA é descrito na seção 4.3.1. Na sequência, são estudadas as variações de parâmetros do *kernel* polinomial (na seção 4.3.2) e do *kernel* gaussiano (na seção 4.3.3) na geração de classificadores com os casos de PSF e PCF do Experimento 2, descrita na seção 2.4.3.

4.3.1 Pré-processamento

De forma equivalente ao descrito na seção 3.4.1, as variáveis de processo são condicionadas para possuir média zero e são normalizadas pelos seus respectivos desvios-padrão. Ressalta-se que, conforme necessário na seleção de variáveis via PFA, as médias e desvios-padrão são calculados utilizando os dados disponíveis na fase de treinamento para cada uma das 20 combinações de grupos. Com isso os ajustes de média e desvio-padrão também são realizados dentro de cada uma destas combinações.

4.3.2 *Kernel* polinomial

Para avaliar o desempenho das características geradas pelas projeções do conjunto de treino utilizando o KPCA com *kernel* polinomial, o expoente de polinômio d foi escolhido variando de 1 até 10, onde o expoente 1 equivale à aplicação do PCA linear conforme apresentado no Capítulo 3.

Como os números de elementos que compõem cada grupo diferem entre si, o número de características obtidas em cada combinação de grupos variou entre 49 e 54. Para cada um dos expoentes d, os valores do F1-*score* médio da classificação dos conjuntos de teste foi calculado.

Considera-se como valor de referência para comparação de resultados o F1-*Score* médio obtido no Experimento 2, de valor igual a 0,7212. Na Figura 4.1, observa-se que todos os resultados foram piores que a referência. O melhor resultado possui valor igual a 0,533, referente ao expoente de valor 1. Este F1-*Score* corresponde a uma taxa de acerto médio de 67,25% para as PSFs e de 62,13% para as PCFs.

Nota-se que houve uma redução substancial na dimensão do conjunto de dados para treinamento e teste dos classificadores. Ao passo que no domínio temporal são utilizadas, no mínimo, 12960 amostras, no domínio transformado a quantidade máxima de pontos disponível é 54. É possível que o número pequeno de casos disponíveis durante a fase de treinamento não seja suficiente para gerar características que tornem possível a separação adequada dos casos de PSF e PCF.

4.3.3 *Kernel* gaussiano

Para utilizar o *kernel* gaussiano, é necessário estudar a região onde o valor de σ produz matrizes $\mathbf{K}$ válidas. Percebe-se que, se o valor for muito pequeno, o expoente tenderá a $-\infty$, gerando uma matriz de zeros. Por outro lado, se o valor de σ for muito grande em relação ao módulo do vetor de diferenças, o expoente tenderá a zero, gerando uma matriz de valores igual a 1. Nenhum destes dois casos são desejáveis para o levantamento de características associadas a relações não lineares

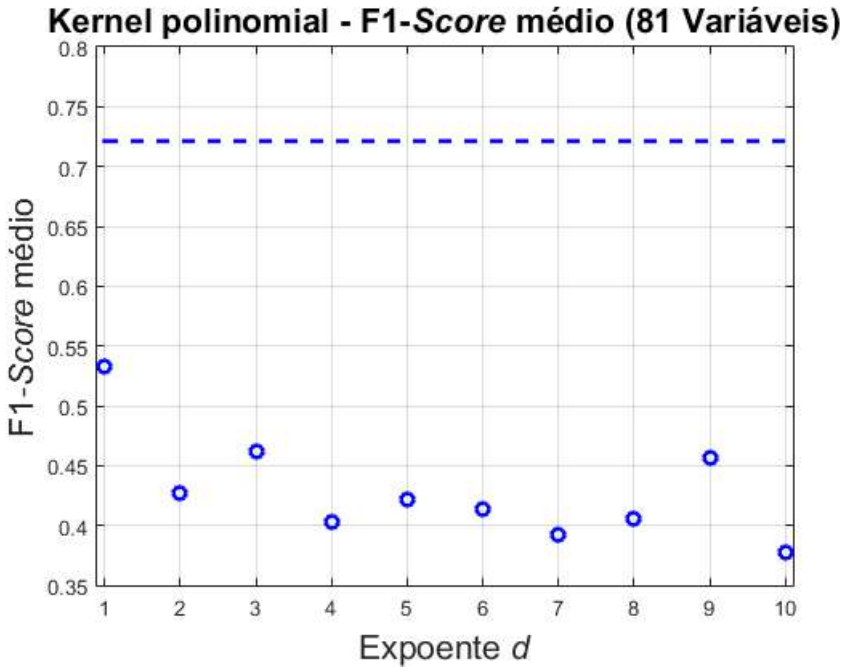


Figura 4.1: Observa-se que todos os resultados utilizando *kernel* polinomial foram piores que a referência. O melhor resultado possui valor igual a 0,533, referente ao expoente de valor 1.

entre as variáveis.

Para definir a melhor região, o Experimento 2 foi realizado variando σ com potências de 10 variando entre 10^{-1} e 10^4. A região que se mostrou mais adequada para levantamento de características foi entre 10^3 e 10^4. Os valores altos justificam-se pelo fato de o tamanho do vetores de diferenças ser igual a 116640 amostras. Dentro desta região, foram obtidos classificadores para 8 pontos contidos no intervalo $]10^3, 10^4[$, separados de 10^3 unidades entre eles.

Novamente, o valor de referência para comparação de resultados o F1-*Score* médio obtido no Experimento 2, de valor igual a 0,7212. Na Figura 4.2, observa-se que todos os resultados foram piores que a referência. O melhor resultado possui valor igual a 0,5025, referente à abscissa σ de valor 3000. Este F1-*Score* corresponde a uma taxa de acerto médio de 73,22% para as PSFs e de 52,95% para as PCFs.

4.4 Conclusão

Este capítulo estudou a aplicabilidade do levantamento de relações não lineares no conjunto de dados dos turbogeradores através de domínios transformados. Conforme [21, 22] a extração de características não lineares é possível aplicando o PCA em um domínio transformado não linearmente. No KPCA, utilizando funções de *kernel*, é possível realizar esta operação sem a necessidade do mapeamento para o domínio transformado de forma explícita, garantindo um menor custo computa-

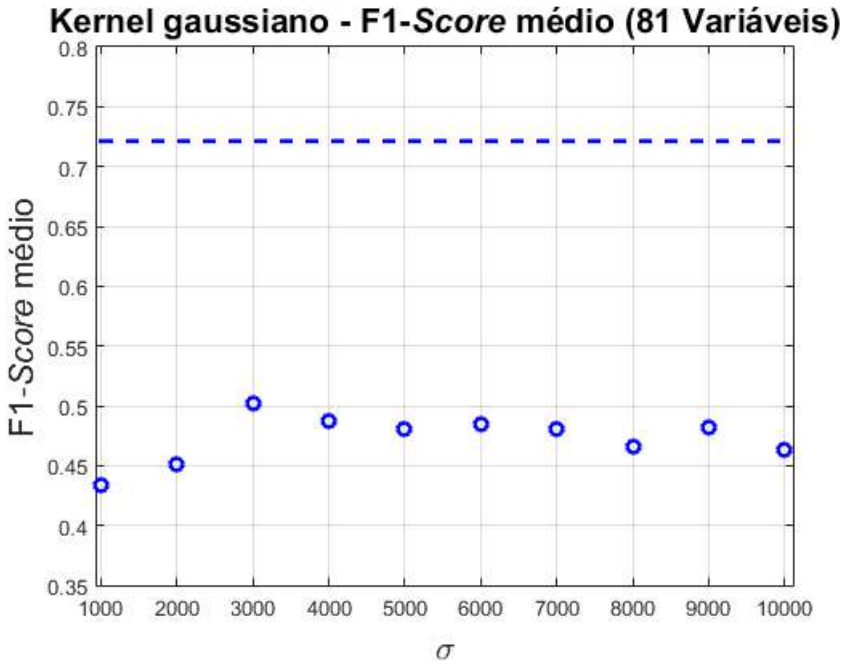


Figura 4.2: Observa-se que todos os resultados foram piores que a referência. O melhor resultado possui valor igual a 0,5025, referente à abscissa σ de valor 3000.

cional. Esta técnica é tradicionalmente utilizada para definir espaços onde sejam ressaltadas novas características que permitam melhorar a separação entre classes.

A aplicação do KPCA foi descrita na seção 4.3, mostrando as etapas de pré-processamento e do estudo da variação dos parâmetros do *kernel* gaussiano e do *kernel* polinomial. Verificou-se que o ajuste de parâmetros das funções de *kernel* é um processo bastante demorado.

Os resultados obtidos não alcançaram o desempenho dos classificadores dos Capítulos 2 e 3. Devido à configuração proposta, o conjunto de características obtidas pelo KPCA possui dimensão muito reduzida, sendo consequência direta de o número de casos de PSF e PCF estudados ser relativamente pequeno.

A expansão do banco de dados e o estudo de novas funções de *kernel* possuem potencial para geração de uma representação que permita a separação mais eficiente dos casos de PCF e PSF estudados.

Capítulo 5

Expansão do Banco de Dados

5.1 Introdução

Embora o banco de dados original, conforme descrito no Capítulo 2, possua o histórico de 3 anos de operação, o número de PCFs e PSFs é relativamente pequeno. Logo, como opção para avaliar a viabilidade de construção de classificadores, conforme [3], apresentado na seção 2.4.3, foi utilizada uma metodologia de validação cruzada. De acordo com os resultados apresentados em [3] e no Capítulo 3, é possível que a geração de um classificador com estes dados seja capaz de obter boa performance ao analisar novos casos de PCFs e PSFs.

Para viabilizar o teste efetivo desta base de dados inicial, foram adquiridos os dados históricos dos 3 anos e meio seguintes (de janeiro 2013 a julho de 2016) das 81 variáveis disponíveis dos mesmos 4 turbogeradores usados no banco inicial. A seção 5.2 apresenta a nova distribuição de PCFs e PSFs detalhadamente.

Na seção 5.3, os casos provenientes das paradas mais recentes são utilizados como dados de teste nos classificadores que obtiveram melhor resultado na seção 3.4.3. A seção 5.3.1 investiga os efeitos da inserção dos novos casos de PCF e PSF nos grupos de casos de parada do Experimento 2, utilizando as variáveis selecionadas com melhor desempenho da seção 3.4.3 para construção de grupos de classificadores. Já na seção 5.4, todo o processo de seleção de variáveis através da aplicação do PFA, conforme seção 3.4.3, é aplicado para criação de novos conjuntos de classificadores, que são comparados com os resultados obtidos na seção anterior. A seção 5.5 emprega a técnica de seleção de PSFs, apresentada em [3], para melhorar o desempenho de classificação. Por fim, a seção 5.6 sintetiza os estudos realizados e indica as principais conclusões das análises propostas.

5.2 Novos dados adquiridos

Novos dados, referentes ao período de janeiro de 2013 a julho de 2016, foram adquiridos via extração das variáveis do banco de dados do historiador, mantendo a taxa de amostragem de 1 amostra por minuto. Estes dados foram tratados seguindo todas as etapas de pré-processamento descritas na seção 2.4.

Consultando o banco de dados de manutenção anotados, buscaram-se todas ocorrências de PSFs que possuem 24 horas de operação ininterrupta precedendo a parada. Para os casos de PCF, todas as falhas priorizadas apresentadas na Tabela 2.1 com 24 horas de operação foram tratadas. Com isso um total de 27 PCFs e 330 PSFs foram identificadas. A Tabela 5.1 apresenta como os casos de PCF e PSF destes novos dados se distribui entre os turbogeradores.

Tabela 5.1: Ocorrências de PSFs e PCFs de janeiro de 2013 até julho de 2016.

Evento	TGA	TGB	TGC	TGD
OIF	0	0	0	0
GDF	4	0	1	2
GUF	1	0	0	1
LUF	1	1	0	0
IGF	0	0	0	0
ETH	0	2	0	0
RLT	0	0	0	0
DLF	0	4	2	0
VPE	2	1	1	4
PSF	80	82	79	89

Percebe-se que o número de PCFs total diminuiu e diversificou-se. A OIF, que possuía maior frequência, não ocorreu nos 3 anos e meio seguintes. A GUF, que era a segunda mais frequente, foi identificada somente em 2 ocasiões. A GDF possui 7 novas ocorrências em relação a somente 4 casos no período de 2010 a 2012.

5.3 Utilização como conjunto de testes

Com a extensão da base de dados, é possível avaliar se os modelos de classificação criados com os dados anteriores são capazes de classificar corretamente os novos casos de PSFs e PCFs. O primeiro teste realizado consistiu na classificação pelo conjunto de classificadores de melhor desempenho treinados na seção 3.4.3, que utilizou variáveis selecionadas pelo PFA modelado pelo grupo de PSFs, com 81 variáveis, sem seleção de variáveis adicionais. O F1-*Score* médio destes classificadores obtido foi de 0,0847, que corresponde a uma taxa de acertos média de 17,67% para PCFs e de 74,03% para PSFs.

Para o segundo teste, seguiu-se [12], que indica o uso de metodologias de validação cruzada para a definição de parâmetros utilizados para determinar o classificador que será treinado com todos os dados disponíveis. Desta forma, utilizando todas as PSFs para modelagem e adotando retenção de 90% de variabilidade, selecionaram-se através do PFA as variáveis, das 81 disponíveis, para treinar o classificador. A quantidade de variáveis selecionadas pelo PFA, considerando somente o número de CPs mínimo, foi igual a 19.

Foram realizadas 5 repetições de treinamento de classificadores. Utilizando estes classificadores, o F1-*Score* médio possui o valor de 0,0263, correspondente a uma taxa de acertos média de 2,22% para PCFs e de 85,45% para PSFs. A única PCF que foi corretamente identificada por 3 dos 5 classificadores treinados foi do tipo VPE da máquina TGD. Estes resultados sugerem que a quantidade de casos de PCF referentes ao período de janeiro de 2010 a dezembro de 2012 não foi suficiente para modelar de forma adequada os casos de PCF. Com isso, para avaliação do método de seleção de variáveis proposto, novas modelagens são estudadas utilizando o conjunto completo de PCFs nas seções 5.3.1 e 5.4.

5.3.1 Desenvolvimento de novos classificadores

Tendo em vista que a base de dados antiga não foi capaz de gerar classificadores capazes de identificar, com bom desempenho, os novos casos de PCF, o passo seguinte foi a construção de novos classificadores utilizando a base de dados estendida.

Com o intuito de estudar a influência isolada de cada tipo de parada adicionada, os casos de PCF e PSF do Experimento 2 mantêm sua distribuição equivalente nos 6 grupos. Com esta estratégia, a seleção de variáveis realizada por combinação de grupos, na seção 3.4.3, pôde ser utilizada.

Os novos casos de PCFs e PSFs foram ordenados por data de ocorrência e por máquina afetada. Seguiu-se com a distribuição pelos 6 grupos em cinco estratégias diferentes:

1. Experimento 3: somente as novas PCFs classificadas como GDF e GUF, que são casos equivalentes aos do banco de dados original;

2. Experimento 4: somente as novas ocorrências de PCFs, de todos tipos identificados;

3. Experimento 5: somente as novas ocorrências de PSFs;

4. Experimento 6: somente as novas ocorrências de PSFs em conjunto com as novas PCFs classificadas como GDF e GUF;

5. Experimento 7: todos os novos casos de PCF e PSF.

• **Experimento 3: novas GDFs e GUFs**

Os 7 novos casos de GDF e os 2 novos casos de GUF foram adicionados aos grupos existentes, conforme listado na Tabela 5.2. A distribuição dos outros tipos de paradas são apresentadas na Tabela 2.6. Com a associação destas tabelas os 6 grupos possuem as seguintes composições:

- 4 grupos com 12 PSFs e 7 PCFs;
- 2 grupos com 11 PSFs e 7 PCFs;

Tabela 5.2: Distribuição de Novas PCFs (GDF e GUF) por Grupos.

	G1	**G2**	**G3**	**G4**	**G5**	**G6**
TGA-GDF	1	1	0	0	1	1
TGA-GUF	0	0	1	0	0	0
TGC-GDF	0	0	0	1	0	0
TGD-GDF	0	0	1	0	0	1
TGD-GUF	0	0	0	0	1	0

Foram gerados 100 classificadores, referentes a 5 repetições de treinamentos com as 20 combinações de 3 grupos, de forma equivalente aos outros experimentos realizados. O F1-*Score* médio destes classificadores para classificação dos respectivos conjuntos de teste foi 0,6568. Este valor corresponde a uma taxa de acerto médio de 74,88% para as PSFs e de 69,14% para as PCFs.

A Figura 5.3(a) mostra a quantidade de erros de predição por caso de PCF, separados por seu tipo, quando presentes no conjunto de testes, do conjunto de classificadores. Observa-se que o perfil de identificação dos casos de GDF e GUF se assemelha bastante ao mostrado na Figura 3.5(b), referente ao experimento realizado na seção 3.4.3 com 81 variáveis, com ligeiro aumento nos erros de predição nos casos de pior desempenho. O perfil de erros para as OIFs modificou-se, acentuando a distribuição dos erros e diminuindo a sua média. O conjunto de novas GDFs apresentou muitos erros de predição, possuindo somente 1 caso com taxa de erro abaixo de 50%. Já as duas novas GUFs também possuem identificações incorretas, sendo uma delas com taxa de erro superior a 50%.

• **Experimento 4: novas PCFs**

Os 27 novos casos de PCF foram adicionados aos grupos existentes, conforme listado na Tabela 5.3. A distribuição dos outros tipos de paradas são apresentadas na Tabela 2.6. Com a associação destas tabelas os 6 grupos possuem as seguintes composições:

- 4 grupos com 12 PSFs e 10 PCFs;

- 2 grupos com 11 PSFs e 10 PCFs;

Tabela 5.3: Distribuição de Novas PCFs (todas) por Grupos.

	G1	**G2**	**G3**	**G4**	**G5**	**G6**
TGA-GDF	1	1	0	0	0	2
TGA-GUF	0	0	1	0	0	0
TGA-LUF	0	0	0	0	1	0
TGA-VPE	0	0	1	1	0	0
TGB-DLF	0	2	0	0	1	1
TGB-ETH	0	0	0	0	1	0
TGB-LUF	0	0	0	1	0	0
TGB-VPE	0	0	0	0	1	0
TGC-GDF	0	0	0	1	0	0
TGC-DLF	0	0	1	0	0	1
TGC-VPE	1	0	0	0	0	0
TGD-GDF	0	0	1	0	1	0
TGD-GUF	0	0	0	0	1	0
TGD-VPE	1	1	0	1	0	1

Foram gerados 100 classificadores, referentes a 5 repetições de treinamentos com as 20 combinações de 3 grupos, de forma equivalente aos outros experimentos realizados. O F1-*Score* médio destes classificadores para classificação dos respectivos conjuntos de teste foi 0,6247. Este valor corresponde a uma taxa de acerto médio de 69,18% para as PSFs e de 62,27% para as PCFs.

A Figura 5.6(a) mostra a quantidade de erros de predição por caso de PCF, separados por seu tipo, quando presentes no conjunto de testes, do conjunto de classificadores. Observa-se que a adição de novos tipos de PCFs degradou a identificação dos casos de GDF, GUF e OIF em comparação com a Figura 3.5(b), referente ao experimento realizado na seção 3.4.3 com 81 variáveis. O conjunto de novas GDFs apresentou menos erros de predição que o Experimento 3, possuindo 3 casos com taxa de erro abaixo de 50%. Já as duas novas GUFs também possuem identificações incorretas, com os erros mais distribuídos entre os dois casos, diferentemente do Experimento 3. Dos novos tipos de PCF, 2 casos de DLF, 1 de ETH, 1 de LUF e 2 de VPE são identificados incorretamente em mais de 50% dos oportunidades em que foram testados.

• Experimento 5: novas PSFs

Os 330 novos casos de PSF foram adicionados aos grupos existentes, conforme listado na Tabela 5.4. A distribuição dos outros tipos de paradas são apresentadas na Tabela 2.6. Com a associação destas tabelas os 6 grupos possuem as seguintes composições:

- 3 grupos com 67 PSFs e 6 PCFs;
- 1 grupo com 67 PSFs e 5 PCFs;
- 2 grupos com 66 PSFs e 5 PCFs;

Tabela 5.4: Distribuição de Novas PSFs por Grupos.

	G1	**G2**	**G3**	**G4**	**G5**	**G6**
TGA-PSF	13	13	13	13	14	14
TGB-PSF	14	14	14	14	13	13
TGC-PSF	13	13	13	13	13	14
TGD-PSF	15	15	15	15	15	14

Foram gerados 100 classificadores, referentes a 5 repetições de treinamentos com as 20 combinações de 3 grupos, de forma equivalente aos outros experimentos realizados. O F1-*Score* médio destes classificadores para classificação dos respectivos conjuntos de teste foi 0,6112. Este valor corresponde a uma taxa de acerto médio de 92,21% para as PSFs e de 85,05% para as PCFs.

A Figura 5.9(a) mostra a quantidade de erros de predição por caso de PCF, separados por seu tipo, quando presentes no conjunto de testes, do conjunto de classificadores. Observa-se que o perfil de identificação dos casos de GDF, GUF e OIF é melhor que o mostrado na Figura 3.5(b), referente ao experimento realizado na seção 3.4.3 com 81 variáveis. Verifica-se que ainda há 1 GUF e 2 OIFs com taxa de erro maior que 50%.

• Experimento 6: novas PSFs em conjunto com novas GDFs e GUFs

Os 330 novos casos de PSF, os 7 novos casos de GDF e os 2 novos casos de GUF foram adicionados aos grupos existentes, conforme listado nas Tabela 5.2 e 5.4. A distribuição dos outros tipos de paradas são apresentadas na Tabela 2.6. Com a associação destas tabelas os 6 grupos possuem as seguintes composições:

- 4 grupos com 67 PSFs e 7 PCFs;
- 2 grupos com 66 PSFs e 7 PCFs;

Foram gerados 100 classificadores, referentes a 5 repetições de treinamentos com as 20 combinações de 3 grupos, de forma equivalente aos outros experimentos realizados. O F1-*Score* médio destes classificadores para classificação dos respectivos conjuntos de teste foi 0,4993. Este valor corresponde a uma taxa de acerto médio de 87,54% para as PSFs e de 71,76% para as PCFs.

A Figura 5.12(a) mostra a quantidade de erros de predição por caso de PCF, separados por seu tipo, quando presentes no conjunto de testes, do conjunto de classificadores. Observa-se que o perfil de identificação dos casos de GDF, GUF e OIF é melhor que o mostrado na Figura 3.5(b), referente ao experimento realizado na seção 3.4.3 com 81 variáveis. Verifica-se que ainda há 2 OIFs com taxa de erro maior que 50%. O conjunto de novas GDFs e GUFs apresentou muitos erros de predição, onde todos os casos possuem taxa de erro acima de 50%.

• Experimento 7: novas PSFs e PCFs

Os 330 novos casos de PSF e os todos os novos casos de PCF foram adicionados aos grupos existentes, conforme listado na Tabela 5.3 e 5.4. A distribuição dos outros tipos de paradas são apresentadas na Tabela 2.6. Com a associação destas tabelas os 6 grupos possuem as seguintes composições:

- 4 grupos com 67 PSFs e 10 PCFs;
- 2 grupos com 66 PSFs e 10 PCFs;

Foram gerados 100 classificadores, referentes a 5 repetições de treinamentos com as 20 combinações de 3 grupos, de forma equivalente aos outros experimentos realizados. O F1-*Score* médio destes classificadores para classificação dos respectivos conjuntos de teste foi 0,4726. Este valor corresponde a uma taxa de acerto médio de 81,95% para as PSFs e de 67,67% para as PCFs.

A Figura 5.15(a) mostra a quantidade de erros de predição por caso de PCF, separados por seu tipo, quando presentes no conjunto de testes do conjunto de classificadores. Observa-se que o perfil de identificação dos casos de GDF, GUF e OIF é melhor que o mostrado na Figura 3.5(b), referente ao experimento realizado na seção 3.4.3 com 81 variáveis. Verifica-se que ainda há 2 OIFs e 1 GUF com taxa de erro maior que 50%. O conjunto de novas GDFs e GUFs apresentou muitos erros de predição, onde 4 casos de GDF e 1 de GUF possuem taxa de erro acima de 50%. Dos novos tipos de PCF, 3 casos de DLF, e de ETH, 2 de LUF, 4 de VPE são identificados incorretamente em mais de 50% dos oportunidades em que foram testados.

• Análise conjunta dos novos experimentos

Após execução dos novos experimentos, verifica-se, novamente, que a seleção de variáveis realizada no Capítulo 3 não é ideal para representação dos novos casos de PCF, resultando no aumento da quantidade de erros de predição desta classe. Por outro lado, ao analisar os resultados do Experimento 5 e ao comparar o Experimento 3 com o Experimento 6, assim como quando se compara o Experimento 4 com o

Experimento 7, percebe-se que a adição dos novos casos de PSF proporcionou uma melhoria na identificação das PCFs em geral.

5.4 Seleção de variáveis com a nova base de dados

Nesta seção, a metodologia de seleção de variáveis proposta na seção 3.4.3 é aplicada novamente para cada uma das distribuições de novos casos de PCF e PSF propostas na seção 5.4. Estas distribuições diferem das distribuições anteriores, uma vez que para seguir a metodologia descrita na seção 2.4.3, é necessária a ordenação temporal das novas PSFs e PCFs.

As estatísticas das variáveis foram modeladas através dos grupos de PSFs, de PCFs e do agrupamento total de PCFs e PSFs, utilizando o conjunto de 81 variáveis definido na seção 3.4.1. O número de variáveis selecionadas variou entre a quantidade de CPs identificada, para dada combinação, e a mesma acrescida de até 5 variáveis adicionais, conforme proposto em [14].

Para cada uma das distribuições de novos casos de PSF e PCF, geraram-se 18 blocos de 20 conjuntos de variáveis para cada combinação de modelagem estatística e número de variáveis selecionadas. Nas subseções seguintes são apresentados os resultados da execução destes experimentos.

5.4.1 Experimento 3: novas GDFs e GUFs

O Experimento 3 foi realizado novamente com todos os 18 blocos de conjuntos de variáveis obtidos pela combinação das 3 modelagens estatísticas propostas. Após a ordenação temporal dos novos casos de GUF e GDF, a nova distribuição dos casos de PCF é mostrada na Tabela 5.5. A distribuição das PSFs ainda segue o original do Experimento 2, conforme Tabela 2.6.

A comparação de desempenho dos blocos de classificadores é realizada através do valor da média do F1-*Score* obtido nas 5 execuções das 20 combinações dos 6 grupos de PSFs e PCFs.

Na Figura 5.1, observa-se que nenhum dos blocos de conjuntos de variáveis alcançou resultados melhores que o Experimento 3 original utilizando os conjuntos de variáveis da seção 3.4.3. O bloco de melhor desempenho é composto pelos conjuntos de variáveis, com 5 variáveis adicionais, selecionados através da modelagem estatística a partir de casos de PSFs e PCFs em conjunto. A quantidade total de variáveis destes conjuntos variou entre 19 e 20. O valor do F1-*Score* médio deste conjunto é $0,6443$. Este valor corresponde a uma taxa de acerto médio de 76,77% para as PSFs e de 65,81% para as PCFs.

Tabela 5.5: Distribuição de PCFs, incluindo novas GDF e GUF, por grupos para nova seleção de variáveis.

	G1	G2	G3	G4	G5	G6
TGA-OIF	0	0	1	0	1	0
TGA-GDF	2	1	1	2	0	1
TGA-GUF	0	1	0	0	0	1
TGB-OIF	1	0	0	0	0	0
TGB-GDF	0	0	0	0	1	0
TGB-GUF	1	2	1	1	1	1
TGC-OIF	1	1	2	2	2	2
TGC-GDF	0	1	0	0	0	0
TGD-OIF	2	1	1	1	0	0
TGD-GDF	0	0	1	0	0	1
TGD-GUF	0	0	0	1	2	1

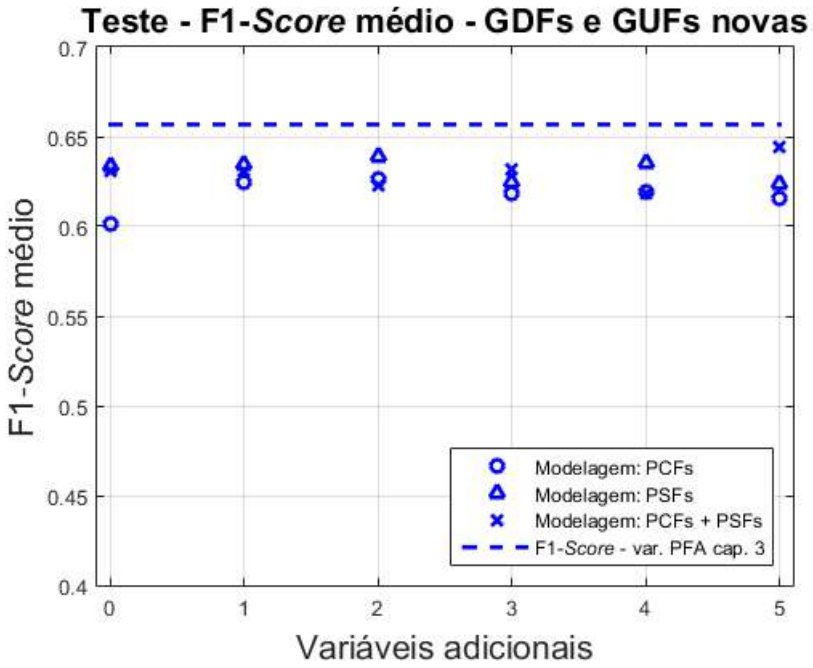


Figura 5.1: Com a seleção de variáveis adicionando somente novos casos de GDFs e GUFs, observa-se que nenhum bloco de conjuntos de variáveis alcançou resultados melhores que o Experimento 3 com as variáveis selecionadas na seção 3.4.3. O F1-*Score* médio do conjunto de melhor desempenho é $0,6443$.

Verificou-se que as variáveis selecionadas variaram bastante para cada combinação de grupos. A quantidade de combinações em que cada uma das variáveis foi selecionada é apresentada na Figura 5.2. A Tabela 5.6 lista as 11 variáveis presentes em mais de 50% das combinações de grupos. Neste bloco, 25 variáveis, listadas na Tabela 5.7, não foram selecionadas em nenhuma das combinações.

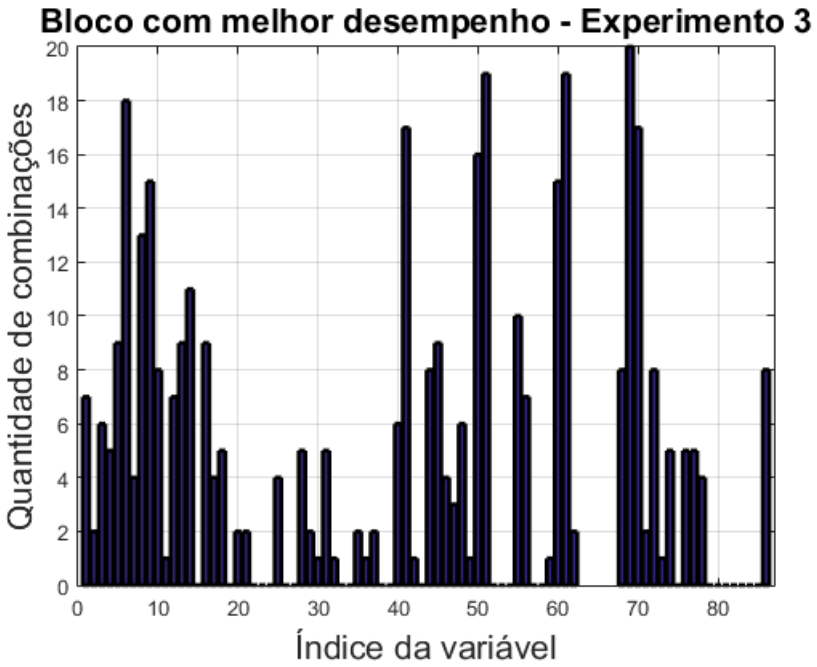


Figura 5.2: No bloco de conjuntos de variáveis com melhor desempenho médio, observa-se que 11 variáveis, listadas na Tabela 5.6, estão presentes em mais de 10 combinações de grupos. Há, também, 25 variáveis, listadas na Tabela 5.7, não selecionadas em nenhuma das combinações.

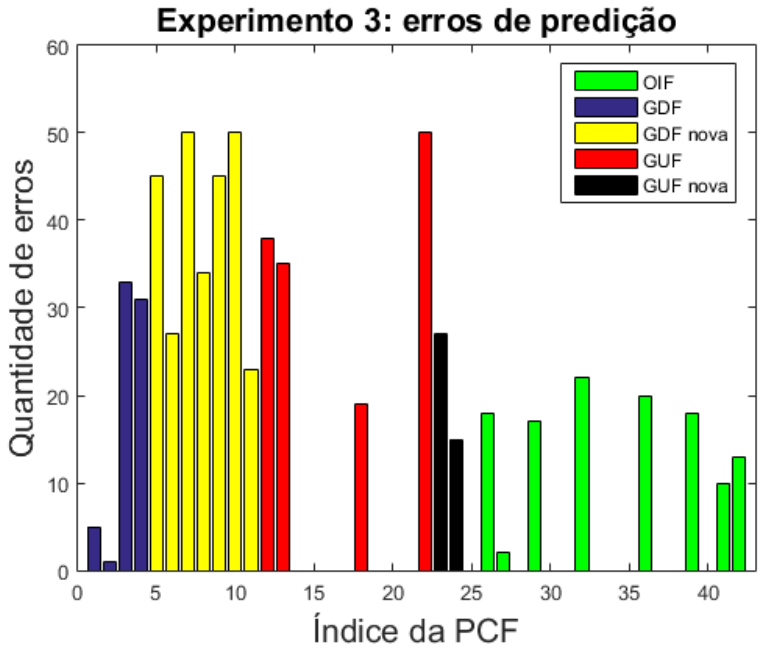
Tabela 5.6: PFA, Experimento 3: variáveis presentes em mais de 50% das combinações.

Índice	**Identificação**	**Descrição**	**Combinações**
6	TT-001	Temperatura ambiente	18
8	TT-003	Temperatura do *manifold* de gás combustível	13
9	VE-001	Vibração na entrada do GG	15
14	VE-006	Vib. PT acopl. X	11
41	PDT-002	Pressão diferencial - Filtro de gás combustível	17
50	ST-002	Rotação Motor de Arranque	16
51	ST-003	Rot. Com. / Turb. Alta	19
60	JY-001	Energia Real	15
61	JQ-001	*Power Factor*	19
69	TI-007	Gerador temp. enrolamento L1-3	20
70	TI-008	Exaustão GG	17

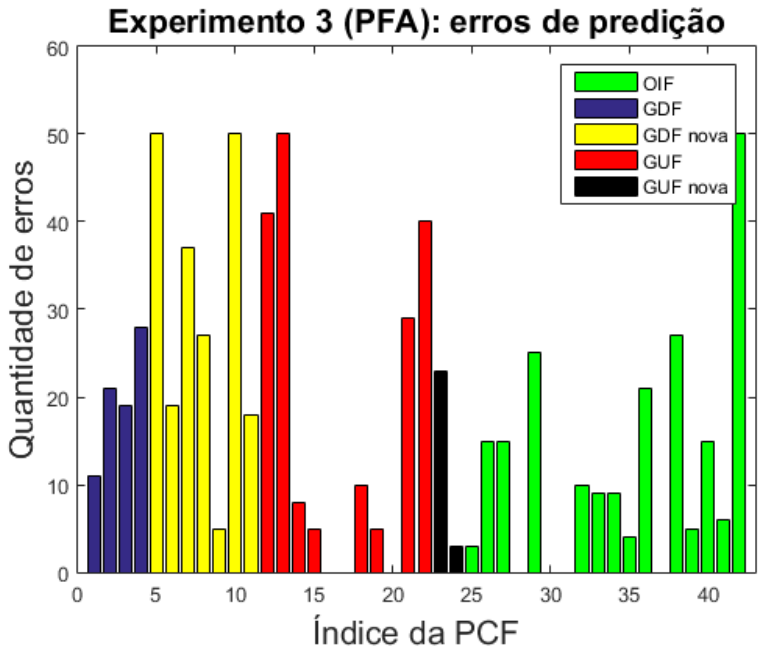
Tabela 5.7: PFA, Experimento 3: variáveis não selecionadas nas combinações.

Índice	**Identificação**	**Descrição**
15	VE-007	Vib. PT acopl. Y
19	ZE-004	*Gearbox* HSS Axial
22	TI-003-03	Termopar 03 do perfil da exaustão
23	TI-003-04	Termopar 04 do perfil da exaustão
24	TI-003-05	Termopar 05 do perfil da exaustão
26	TI-003-07	Termopar 07 do perfil da exaustão
27	TI-003-08	Termopar 08 do perfil da exaustão
33	TI-003-13	Termopar 13 do perfil da exaustão
34	TI-003-14	Termopar 14 do perfil da exaustão
38	TI-005	Gerador temp. enrolamento L2-2
39	TI-006	Gerador temp. enrolamento L2-1
43	PI-001	Pressão *header* óleo mineral
52	ST-004	Rotação PT
53	TE-001	Exaustão GG
54	IT-001	Corrente de excitação do campo
57	ET-001	Voltagem ca
58	JT-001	Potência Reativa
75	TI-013	*Gearbox* LSS Temperatura mancal DE
79	TI-017	Gerador ar de resfriamento (quente) DE
80	TI-018	Gerador ar de resfriamento (quente) NDE
81	TI-019	Gerador ar de resfriamento (frio) NDE
82	TI-020	Gerador Temperatura mancal DE
83	TI-021	Gerador temp. enrolamento L1-1
84	TI-022	Gerador temp. enrolamento L1-2
85	TI-023	Gerador temp. enrolamento L3-1

A Figura 5.3(b) mostra a quantidade de erros de predição por caso de PCF, separados por seu tipo, quando presentes no conjunto de testes, referentes ao bloco de conjunto de variáveis com melhor desempenho médio. Observa-se que, em relação aos casos de PCF originais, 1 caso de GUF e 1 de OIF foram identificados incorretamente em todos classificadores treinados. Além disso, verifica-se que um total de 2 GDFs, 3 GUFs e 1 OIF destas PCFs são identificadas incorretamente em mais de 50% das oportunidades em que foram testadas. Das novas GDFs e GUFs, 3 casos de GDF possuem taxa de erro acima de 50%. Em comparação com a Figura 5.3(a), nota-se que, após a aplicação do PFA, os classificadores treinados foram capazes de melhorar seu desempenho em relação ao Experimento 3 com as variáveis selecionadas na seção 3.4.3 para os novos casos de GDF e GUF. No entanto, a capacidade de identificação dos 33 casos de PCF originais degradou consideravelmente, resultando num desempenho global pior, conforme observado na Figura 5.1.



(a)



(b)

Figura 5.3: Erros de predição em casos de PCF no Experimento 3: (a) observa-se que o perfil de identificação dos casos de GDF e GUF se assemelha bastante com o mostrado na Figura 3.5(b), referente ao experimento realizado na seção 3.4.3 com 81 variáveis. O perfil de erros para as OIFs modificou-se, acentuando a distribuição dos erros e diminuindo a sua média. O conjunto de novas GDFs apresentou muitos erros de predição. Já as duas novas GUFs também possuem identificações incorretas; (b) em comparação com (a), nota-se que, após a aplicação do PFA, os classificadores treinados foram capazes de melhorar seu desempenho em relação ao Experimento 3 com as variáveis selecionadas na seção 3.4.3 para os novos casos de GDF e GUF. No entanto, a capacidade de identificação dos 33 casos de PCF originais degradou consideravelmente, resultando num desempenho global pior.

5.4.2 Experimento 4: novas PCFs

O Experimento 4 foi realizado novamente com todos os 18 blocos de conjuntos de variáveis obtidos pela combinação das 3 modelagens estatísticas propostas. Após a ordenação temporal das novas PCFs, a nova distribuição dos casos de PCF é mostrada na Tabela 5.8. A distribuição das PSFs ainda segue o original do Experimento 2, conforme Tabela 2.6.

A comparação de desempenho dos blocos de classificadores é realizada através do valor da média do F1-*Score* obtido nas 5 execuções das 20 combinações dos 6 grupos de PSFs e PCFs.

Tabela 5.8: Distribuição de PCFs, incluindo todas as novas PCFs, por grupos para nova seleção de variáveis.

	G1	**G2**	**G3**	**G4**	**G5**	**G6**
TGA-OIF	0	0	1	0	1	0
TGA-GDF	3	1	1	1	0	1
TGA-GUF	0	1	0	0	0	1
TGA-LUF	0	0	0	1	0	0
TGA-VPE	0	1	0	0	1	0
TGB-OIF	0	0	0	0	0	1
TGB-GDF	0	0	0	1	0	0
TGB-GUF	1	1	2	1	1	1
TGB-DLF	0	2	0	0	1	1
TGB-ETH	1	0	1	0	0	0
TGB-LUF	0	0	0	1	0	0
TGB-VPE	0	0	0	0	1	0
TGC-OIF	2	1	2	2	1	2
TGC-GDF	0	1	0	0	0	0
TGC-DLF	1	0	0	1	0	0
TGC-VPE	0	0	0	0	1	0
TGD-OIF	0	0	2	1	1	1
TGD-GDF	0	0	1	0	1	0
TGD-GUF	1	1	0	0	1	1
TGD-VPE	1	1	0	1	0	1

Na Figura 5.4, observa-se que 13 blocos de conjuntos de variáveis alcançaram resultados melhores que o Experimento 4 original utilizando as variáveis selecionadas na seção 3.4.3. O bloco de melhor desempenho é composto pelos conjuntos de variáveis, com 2 variáveis adicionais, selecionados através da modelagem estatística a partir de casos de PSFs. A quantidade total de variáveis destes conjuntos variou entre 13 e 15. O valor do F1-*Score* médio deste conjunto é $0,6446$. Este valor corresponde a uma taxa de acerto médio de 68,96% para as PSFs e de 64,93% para as PCFs.

Verificou-se que as variáveis selecionadas variaram bastante para cada combinação de grupos. A quantidade de combinações em que cada uma das variáveis foi selecionada é apresentada na Figura 5.5. A Tabela 5.9 lista as 4 variáveis presentes em mais de 50% das combinações de grupos. Neste bloco, 24 variáveis, listadas na Tabela 5.10, não foram selecionadas em nenhuma das combinações.

A Figura 5.6(b) mostra a quantidade de erros de predição por caso de PCF, separados por seu tipo, quando presentes no conjunto de testes, referentes ao bloco de conjunto de variáveis com melhor desempenho médio. Observa-se que, em relação ao casos de PCF originais, 3 casos de GDF, 2 de GUF e 2 de OIF são identificados incorretamente em mais de 50% das oportunidades em que foram testados. Das novas GDFs e GUFs, 2 casos de GDF possuem taxa de erro acima de 50%. Nenhum dos novos tipos de PCF possui taxa de erro acima de 50%. Em comparação com a Figura 5.6(a), nota-se que, após a aplicação do PFA, os classificadores treinados foram capazes de melhorar seu desempenho em relação ao Experimento 4 com as variáveis selecionadas na seção 3.4.3, com exceção dos casos de GDF originais. Esta capacidade resultou num desempenho global melhor, conforme observado na Figura 5.4.

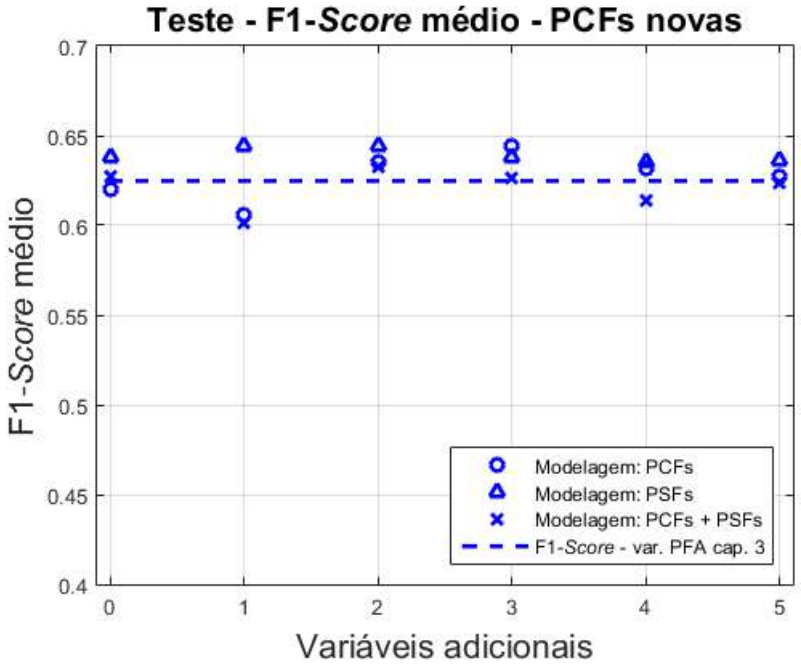


Figura 5.4: Com a seleção de variáveis adicionando todas as novas PCFs, observa-se que 13 blocos de conjuntos de variáveis alcançaram resultados melhores que o Experimento 4 com as variáveis selecionadas na seção 3.4.3. O F1-*Score* médio do conjunto de melhor desempenho é $0,6446$.

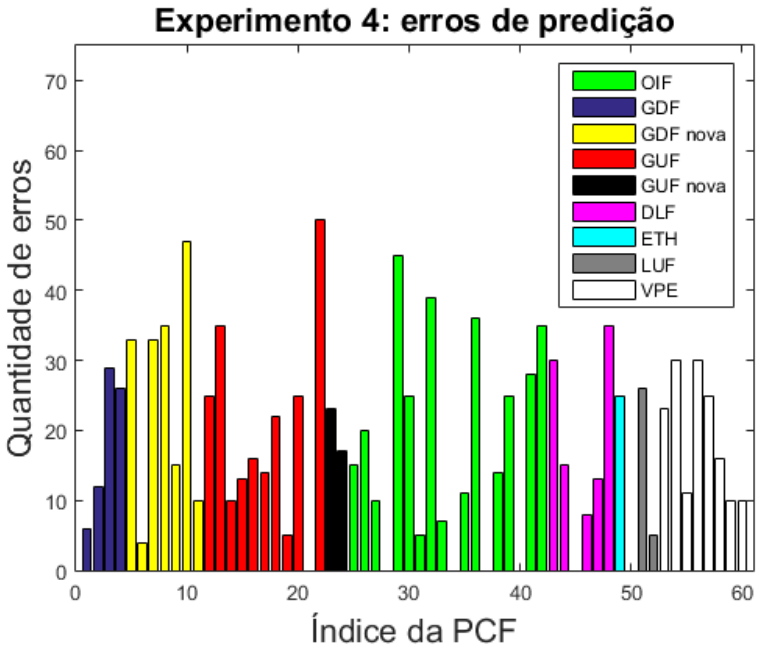
Figura 5.5: No bloco de conjuntos de variáveis com melhor desempenho médio, observa-se que 4 variáveis, listadas na Tabela 5.9, estão presentes em mais de 10 combinações de grupos. Há, também, 24 variáveis, listadas na Tabela 5.7, não selecionadas em nenhuma das combinações.

Tabela 5.9: PFA, Experimento 4: variáveis presentes em mais de 50% das combinações.

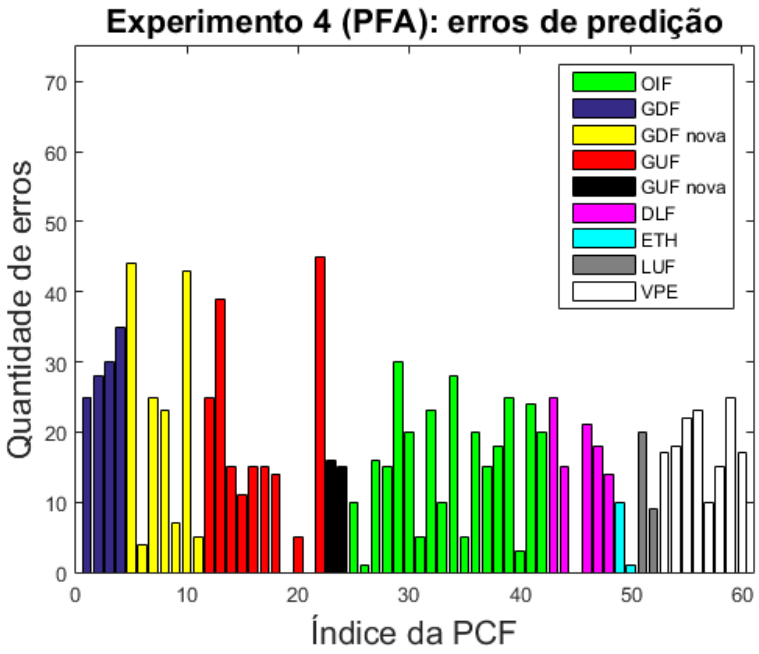
Índice	**Identificação**	**Descrição**	**Combinações**
6	TT-001	Temperatura ambiente	13
9	VE-001	Vibração na entrada do GG	12
69	TI-007	Gerador temp. enrolamento L1-3	18
70	TI-008	Exaustão GG	18

Tabela 5.10: PFA, Experimento 4: variáveis não selecionadas nas combinações.

Índice	**Identificação**	**Descrição**
11	VE-003	Vibração na turbina do GG
21	TI-003-02	Termopar 02 do perfil da exaustão
22	TI-003-03	Termopar 03 do perfil da exaustão
23	TI-003-04	Termopar 04 do perfil da exaustão
24	TI-003-05	Termopar 05 do perfil da exaustão
27	TI-003-08	Termopar 08 do perfil da exaustão
29	TI-003-10	Termopar 10 do perfil da exaustão
30	TI-003-11	Termopar 11 do perfil da exaustão
33	TI-003-13	Termopar 13 do perfil da exaustão
34	TI-003-14	Termopar 14 do perfil da exaustão
35	TI-003-15	Termopar 15 do perfil da exaustão
37	TI-004	Gerador temp. enrolamento L2-3
38	TI-005	Gerador temp. enrolamento L2-2
39	TI-006	Gerador temp. enrolamento L2-1
43	PI-001	Pressão *header* óleo mineral
54	IT-001	Corrente de excitação do campo
58	JT-001	Potência Reativa
76	TI-014	*Gearbox* LSS Temperatura mancal NDE
77	TI-015	*Gearbox* HSS Temperatura mancal NDE
79	TI-017	Gerador ar de resfriamento (quente) DE
80	TI-018	Gerador ar de resfriamento (quente) NDE
81	TI-019	Gerador ar de resfriamento (frio) NDE
83	TI-021	Gerador temp. enrolamento L1-1
84	TI-022	Gerador temp. enrolamento L1-2



(a)



(b)

Figura 5.6: Erros de predição em casos de PCF no Experimento 4: (a) Observa-se que a adição de novos tipos de PCFs degradou a identificação dos casos de GDF, GUF e OIF em comparação com a Figura 3.5(b), referente ao experimento realizado na seção 3.4.3 com 81 variáveis. O conjunto de novas GDFs apresentou menos erros de predição que o Experimento 3. Já as duas novas GUFs também possuem identificações incorretas, com os erros mais distribuídos entre os dois casos, diferentemente do Experimento 3; (b) em comparação com (a), nota-se que, após a aplicação do PFA, os classificadores treinados foram capazes de melhorar seu desempenho em relação ao Experimento 4 com as variáveis selecionadas na seção 3.4.3, com exceção dos casos de GDF originais. Esta capacidade resultou num desempenho global melhor, conforme observado na Figura 5.4.

5.4.3 Experimento 5: novas PSFs

O Experimento 5 foi realizado novamente com todos os 18 blocos de conjuntos de variáveis obtidos pela combinação das 3 modelagens estatísticas propostas. Após a ordenação temporal das novas PSFs, a nova distribuição dos casos de PSF é mostrada na Tabela 5.11. A distribuição das PCFs ainda segue o original do Experimento 2, conforme Tabela 2.6.

A comparação de desempenho dos blocos de classificadores é realizada através do valor da média do F1-*Score* obtido nas 5 execuções das 20 combinações dos 6 grupos de PSFs e PCFs.

Na Figura 5.7, observa-se que 15 blocos de conjuntos de variáveis alcançaram resultados melhores que o Experimento 5 original utilizando as variáveis selecionadas na seção 3.4.3. O bloco de melhor desempenho é composto pelos conjuntos de variáveis, com 4 variáveis adicionais, selecionados através da modelagem estatística a partir de casos de PSFs e PCFs em conjunto. A quantidade total de variáveis destes conjuntos variou entre 18 e 60. O valor do F1-*Score* médio deste conjunto é $0,6703$. Este valor corresponde a uma taxa de acerto médio de 93,59% para as PSFs e de 89,73% para as PCFs.

Verificou-se que as variáveis selecionadas variaram bastante para cada combinação de grupos. A quantidade de combinações em que cada uma das variáveis foi selecionada é apresentada na Figura 5.8. A Tabela 5.12 lista as 28 variáveis presentes em mais de 50% das combinações de grupos. Neste bloco, não houve variáveis não selecionadas nas combinações.

A Figura 5.9(b) mostra a quantidade de erros de predição por caso de PCF, separados por seu tipo, quando presentes no conjunto de testes, referentes ao bloco de conjunto de variáveis com melhor desempenho médio. Observa-se que, em relação ao casos de PCF originais, 1 caso de GUF e 1 de OIF são identificados incorretamente em mais de 50% das oportunidades em que foram testados. Em comparação com a Figura 5.9(a), nota-se que, após a aplicação do PFA, os classificadores treinados foram capazes de melhorar seu desempenho em relação ao Experimento 5 com as variáveis selecionadas na seção 3.4.3, com exceção dos casos de GDF originais. Esta capacidade resultou num desempenho global melhor, conforme observado na Figura 5.7.

Tabela 5.11: Distribuição de PSFs, com inclusão das novas, por grupos.

	G1	**G2**	**G3**	**G4**	**G5**	**G6**
TGA-PSF	17	17	17	17	16	17
TGB-PSF	17	17	16	17	17	16
TGC-PSF	16	16	16	15	16	16
TGD-PSF	17	17	18	18	17	17

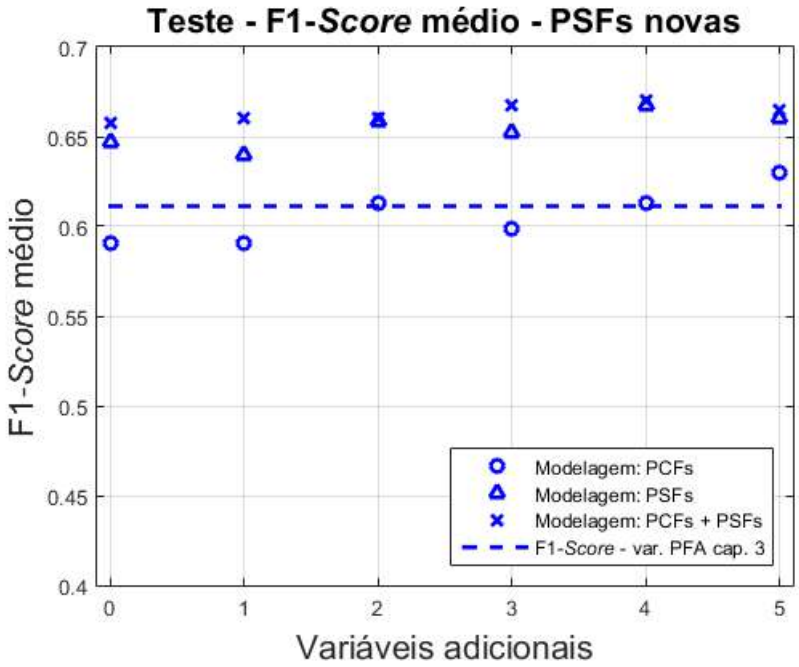


Figura 5.7: Com a seleção de variáveis adicionando todas as novas PSFs, observa-se que 15 blocos de conjuntos de variáveis alcançaram resultados melhores que o Experimento 5 com as variáveis selecionadas na seção 3.4.3. O F1-*Score* médio do conjunto de melhor desempenho é $0,6703$.

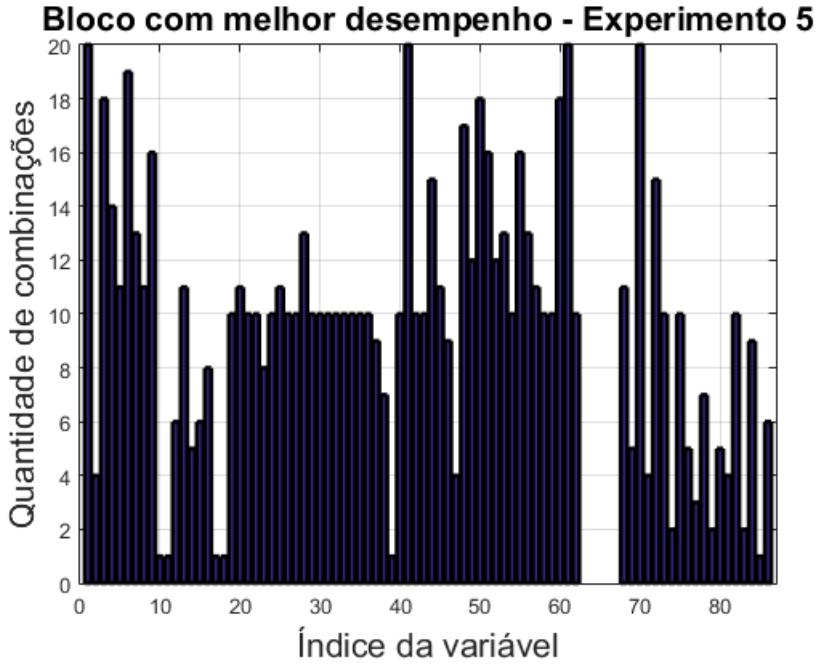
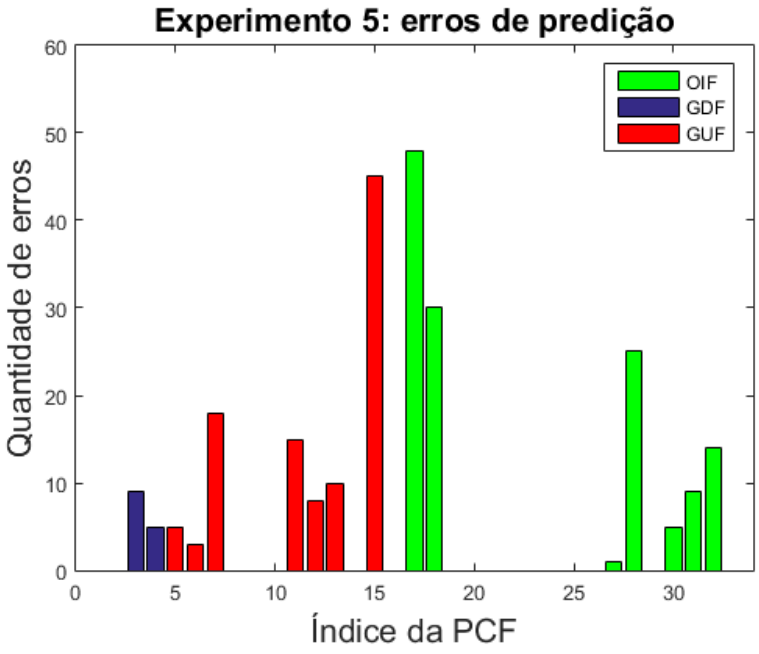


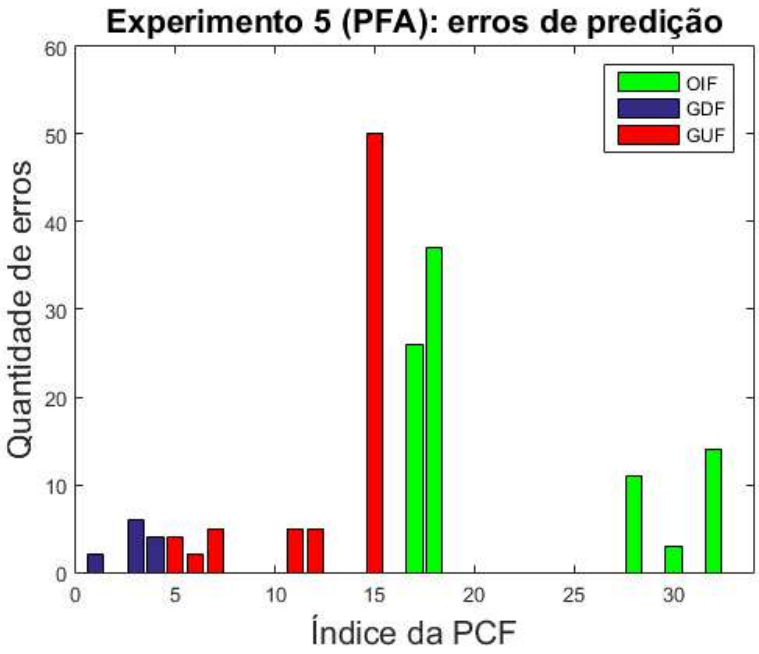
Figura 5.8: No bloco de conjuntos de variáveis com melhor desempenho médio, observa-se que 28 variáveis, listadas na Tabela 5.12, estão presentes em mais de 10 combinações de grupos. Neste bloco, não houve variáveis não selecionadas nas combinações.

Tabela 5.12: PFA, Experimento 5: variáveis presentes em mais de 50% das combinações.

Índice	**Identificação**	**Descrição**	**Combinações**
1	FIT-001	Vazão de gás combustível	20
3	TI-002	Temperatura do *header* de óleo mineral	18
4	TIT-001	Saída de água quente do WHRU	14
5	TIT-002	Temperatura do WHRU	11
6	TT-001	Temperatura ambiente	19
7	TT-002	Temperatura do gás combustível	13
8	TT-003	Temperatura do *manifold* de gás combustível	11
9	VE-001	Vibração na entrada do GG	16
13	VE-005	Vib. PT disc. end. Y	11
20	TI-003-01	Termopar 01 do perfil da exaustão	11
25	TI-003-06	Termopar 06 do perfil da exaustão	11
28	TI-003-09	Termopar 09 do perfil da exaustão	13
41	PDT-002	Pressão diferencial - Filtro de gás combustível	20
44	PIT-001	Pressão	15
45	PT-001	Pressão P1	11
48	PT-003	Gás Comb.	17
49	ST-001	Rot. Com. / Turb. Baixa	12
50	ST-002	Rotação Motor de Arranque	18
51	ST-003	Rot. Com. / Turb. Alta	16
52	ST-004	Rotação PT	12
53	TE-001	Exaustão GG	13
55	ET-001	Voltagem ab	16
56	ET-001	Voltagem bc	13
57	ET-001	Voltagem ca	11
60	JY-001	Energia Real	18
61	JQ-001	*Power Factor*	20
68	IT-002	Corrente	11
70	TI-008	Exaustão GG	20



(a)



(b)

Figura 5.9: Erros de predição em casos de PCF no Experimento 5: (a) Observa-se que o perfil de identificação dos casos de GDF, GUF e OIF é melhor que o mostrado na Figura 3.5(b), referente ao experimento realizado na seção 3.4.3 com 81 variáveis. Verifica-se que ainda há 1 GUF e 2 OIFs com taxa de erro maior que 50%; (b) observa-se que, em relação ao casos de PCF originais, 1 caso de GUF e 1 de OIF são identificados incorretamente em mais de 50% dos oportunidades em que foram testados. Em comparação com (a), nota-se que, após a aplicação do PFA, os classificadores treinados foram capazes de melhorar seu desempenho em relação ao Experimento 5 com as variáveis selecionadas na seção 3.4.3, com exceção dos casos de GDF originais. Esta capacidade resultou num desempenho global melhor, conforme observado na Figura 5.7.

5.4.4 Experimento 6: novas PSFs em conjunto com novas GDFs e GUFs

O Experimento 6 foi realizado novamente com todos os 18 blocos de conjuntos de variáveis obtidos pela combinação das 3 modelagens estatísticas propostas. Após a ordenação temporal das novas GDFs e GUFs, a distribuição dos casos de PCF é mostrada na Tabela 5.5. Com a ordenação temporal das novas PSFs, a distribuição dos casos de PSF é mostrada na Tabela 5.11.

A comparação de desempenho dos blocos de classificadores é realizada através do valor da média do F1-*Score* obtido nas 5 execuções das 20 combinações dos 6 grupos de PSFs e PCFs.

Na Figura 5.10, observa-se que 12 blocos de conjuntos de variáveis alcançaram resultados melhores que o Experimento 6 original utilizando as variáveis selecionadas na seção 3.4.3. O bloco de melhor desempenho é composto pelos conjuntos de variáveis, com 2 variáveis adicionais, selecionados através da modelagem estatística a partir de casos de PSFs e PCFs em conjunto. A quantidade total de variáveis destes conjuntos variou entre 16 e 58. O valor do F1-*Score* médio deste conjunto é $0,5174$. Este valor corresponde a uma taxa de acerto médio de 88,38% para as PSFs e de 73,1% para as PCFs.

Verificou-se que as variáveis selecionadas variaram bastante para cada combinação de grupos. A quantidade de combinações em que cada uma das variáveis foi selecionada é apresentada na Figura 5.11. A Tabela 5.13 lista as 28 variáveis presentes em mais de 50% das combinações de grupos. Neste bloco, 2 variáveis, listadas na Tabela 5.14, não foram selecionadas em nenhuma das combinações.

A Figura 5.12(b) mostra a quantidade de erros de predição por caso de PCF, separados por seu tipo, quando presentes no conjunto de testes, referentes ao bloco de conjunto de variáveis com melhor desempenho médio. Observa-se que, em relação ao casos de PCF originais, 2 casos de OIF destas PCFs são identificadas incorretamente em mais de 50% das oportunidades em que foram testadas. Das novas GDFs e GUFs, somente 1 caso de GDF possui taxa de erro abaixo de 50%. Em comparação com a Figura 5.12(a), nota-se que, após a aplicação do PFA, os classificadores treinados foram capazes de melhorar marginalmente seu desempenho em relação ao Experimento 6 com as variáveis selecionadas na seção 3.4.3 para os novos casos de GDF e GUF. No entanto, a capacidade de identificação dos 33 casos de PCF originais, de forma mais acentuada nos casos de OIF, degradou consideravelmente. Ainda assim, o desempenho global melhorou, conforme observado na Figura 5.10.

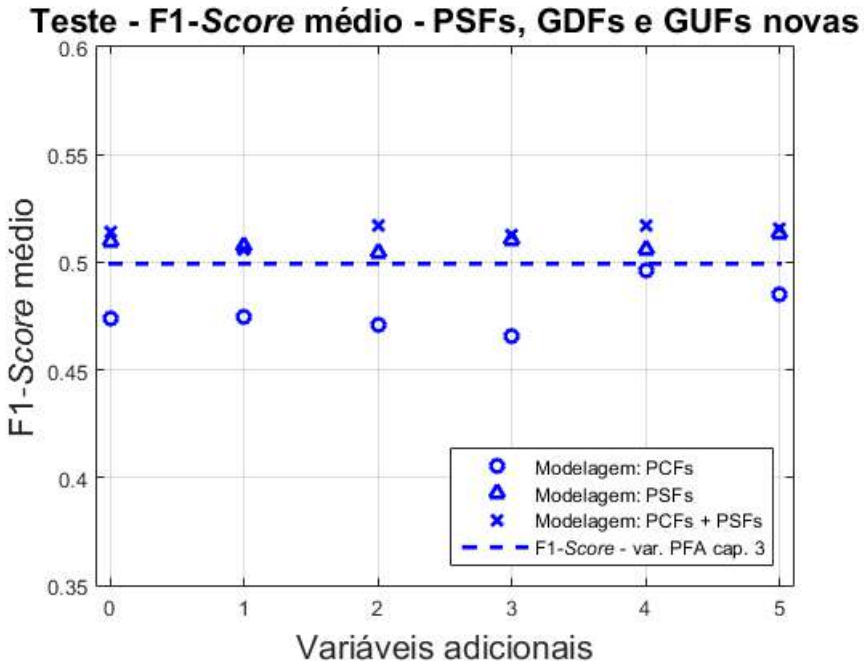


Figura 5.10: Com a seleção de variáveis adicionando as novas GDFs, GUFs e PSFs, observa-se que 12 blocos de conjuntos de variáveis alcançaram resultados melhores que o Experimento 6 com as variáveis selecionadas na seção 3.4.3. O F1-*Score* médio do conjunto de melhor desempenho é $0,5174$.

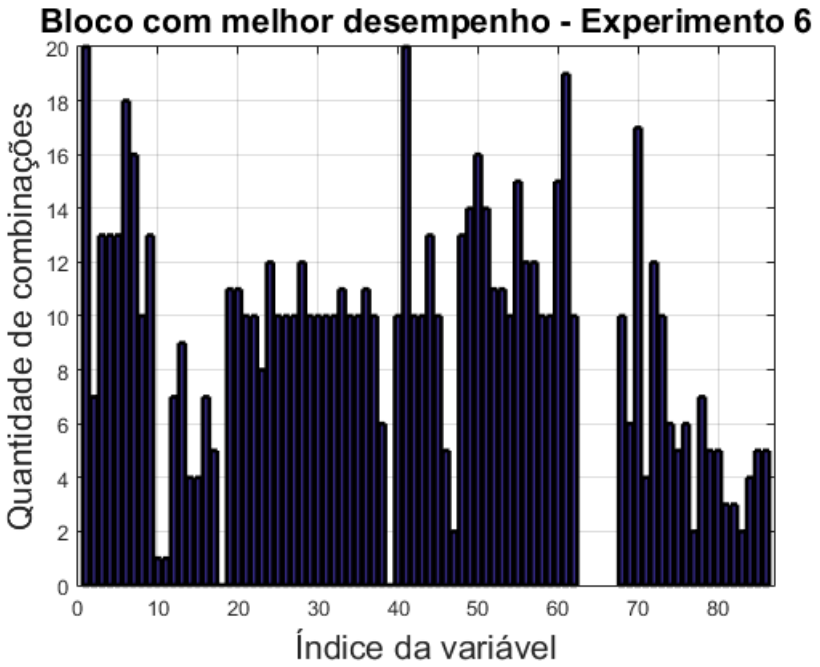


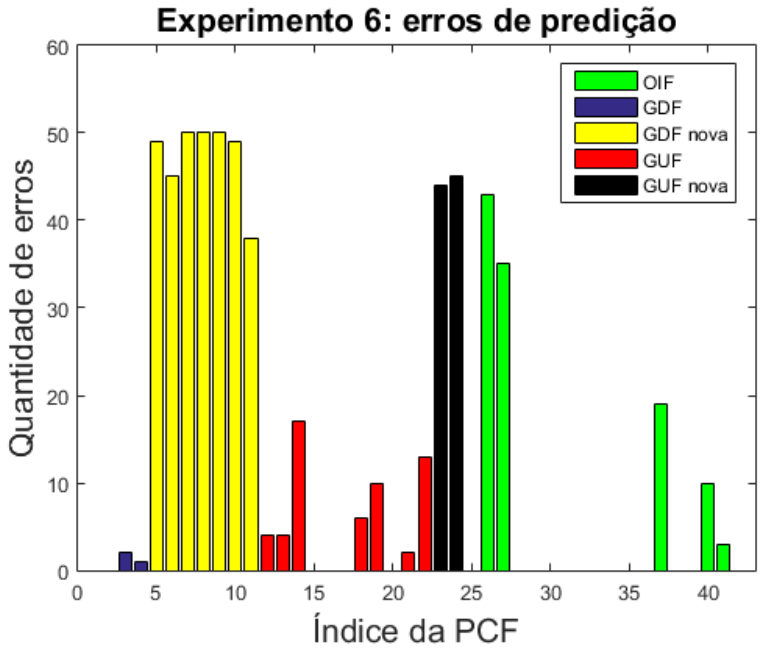
Figura 5.11: No bloco de conjuntos de variáveis com melhor desempenho médio, observa-se que 28 variáveis, listadas na Tabela 5.13, estão presentes em mais de 10 combinações de grupos. Há, também, 2 variáveis, listadas na Tabela 5.14, não selecionadas em nenhuma das combinações.

Tabela 5.13: PFA, Experimento 6: variáveis presentes em mais de 50% das combinações.

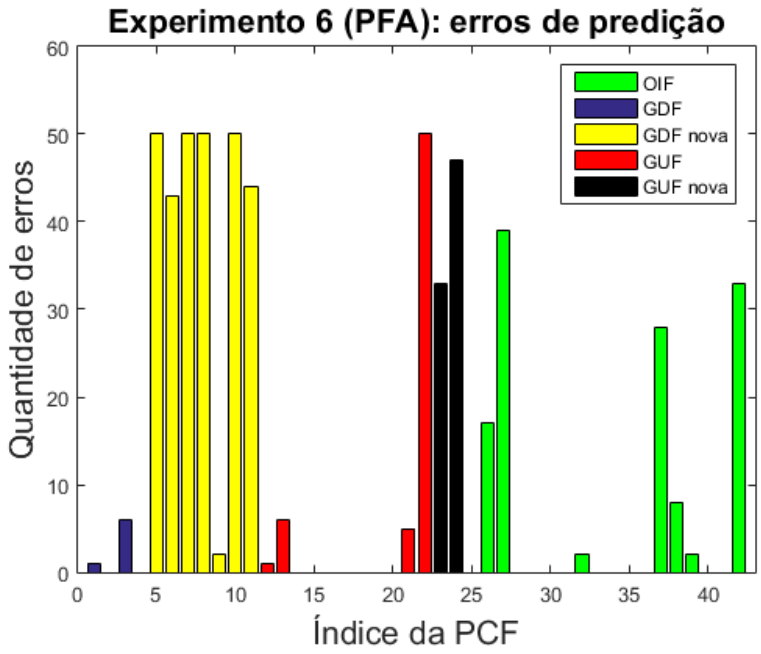
Índice	**Identificação**	**Descrição**	**Combinações**
1	FIT-001	Vazão de gás combustível	20
3	TI-002	Temperatura do *header* de óleo mineral	13
4	TIT-001	Saída de água quente do WHRU	13
5	TIT-002	Temperatura do WHRU	13
6	TT-001	Temperatura ambiente	18
7	TT-002	Temperatura do gás combustível	16
9	VE-001	Vibração na entrada do GG	13
19	ZE-004	*Gearbox* HSS Axial	11
20	TI-003-01	Termopar 01 do perfil da exaustão	11
24	TI-003-05	Termopar 05 do perfil da exaustão	12
28	TI-003-09	Termopar 09 do perfil da exaustão	12
33	TI-003-14	Termopar 14 do perfil da exaustão	11
36	TI-003-17	Termopar 17 do perfil da exaustão	11
41	PDT-002	Pressão diferencial - Filtro de gás combustível	20
44	PIT-001	Pressão	13
48	PT-003	Gás Comb.	13
49	ST-001	Rot. Com. / Turb. Baixa	14
50	ST-002	Rotação Motor de Arranque	16
51	ST-003	Rot. Com. / Turb. Alta	14
52	ST-004	Rotação PT	11
53	TE-001	Exaustão GG	11
55	ET-001	Voltagem ab	15
56	ET-001	Voltagem bc	12
57	ET-001	Voltagem ca	12
60	JY-001	Energia Real	15
61	JQ-001	*Power Factor*	19
70	TI-008	Exaustão GG	17
72	TI-010	PT Temperatura Mancal de Escora	12

Tabela 5.14: PFA, Experimento 6: variáveis não selecionadas nas combinações.

Índice	**Identificação**	**Descrição**
18	ZE-003	*Gearbox* LSS Axial
39	TI-006	Gerador temp. enrolamento L2-1



(a)



(b)

Figura 5.12: Erros de predição em casos de PCF no Experimento 6: (a) observa-se que o perfil de identificação dos casos de GDF, GUF e OIF é melhor que o mostrado na Figura 3.5(b), referente ao experimento realizado na seção 3.4.3 com 81 variáveis. O conjunto de novas GDFs e GUFs apresentou muitos erros de predição, onde todos os casos possuem taxa de erro acima de 50%; (b) em comparação com (a), nota-se que, após a aplicação do PFA, os classificadores treinados foram capazes de melhorar marginalmente seu desempenho em relação ao Experimento 6 com as variáveis selecionadas na seção 3.4.3 para os novos casos de GDF e GUF. No entanto, a capacidade de identificação dos 33 casos de PCF originais, de forma mais acentuada nos casos de OIF, degradou consideravelmente. Ainda assim, o desempenho global melhorou, conforme observado na Figura 5.10.

5.4.5 Experimento 7: novas PSFs e PCFs

O Experimento 7 foi realizado novamente com todos os 18 blocos de conjuntos de variáveis obtidos pela combinação das 3 modelagens estatísticas propostas. Após a ordenação temporal das novas PCFs, a nova distribuição dos casos de PCF é mostrada na Tabela 5.8. Com a ordenação temporal das novas PSFs, a distribuição dos casos de PSF é mostrada na Tabela 5.11.

A comparação de desempenho dos blocos de classificadores é realizada através do valor da média do F1-*Score* obtido nas 5 execuções das 20 combinações dos 6 grupos de PSFs e PCFs.

Na Figura 5.13, observa-se que todos os blocos de conjuntos de variáveis alcançaram resultados melhores que o Experimento 7 original utilizando as variáveis selecionadas na seção 3.4.3. O bloco de melhor desempenho é composto pelos conjuntos de variáveis, com 3 variáveis adicionais, selecionados através da modelagem estatística a partir de casos de PSFs e PCFs em conjunto. A quantidade total de variáveis destes conjuntos variou ente 17 e 59. O valor do F1-*Score* médio deste conjunto é $0,4987$. Este valor corresponde a uma taxa de acerto médio de 83,09% para as PSFs e de 70,08% para as PCFs.

Verificou-se que as variáveis selecionadas variaram bastante para cada combinação de grupos. A quantidade de combinações em que cada uma das variáveis foi selecionada é apresentada na Figura 5.14. A Tabela 5.15 lista as 25 variáveis presentes em mais de 50% das combinações de grupos. Neste bloco, 4 variáveis, listadas na Tabela 5.16, não foram selecionadas em nenhuma das combinações.

A Figura 5.15(b) mostra a quantidade de erros de predição por caso de PCF, separadas por seu tipo, quando presentes no conjunto de testes, referentes ao bloco de conjunto de variáveis com melhor desempenho médio. Observa-se que, em relação ao casos de PCF originais, 1 caso de GUF e 2 de OIF são identificados incorretamente em mais de 50% das oportunidades em que foram testados. Das novas GDFs e GUFs, 3 casos de GDF possuem taxa de erro acima de 50%. Dos novos tipos de PCF, 3 casos de DLF e 5 de VPE possuem taxa de erro acima de 50%. Em comparação com a Figura 5.15(a), nota-se que, após a aplicação do PFA, os classificadores treinados foram capazes de melhorar seu desempenho em relação ao Experimento 7 com as variáveis selecionadas na seção 3.4.3, com exceção dos novos casos de GDF e dos casos de VPE. Esta capacidade resultou num desempenho global melhor, conforme observado na Figura 5.13.

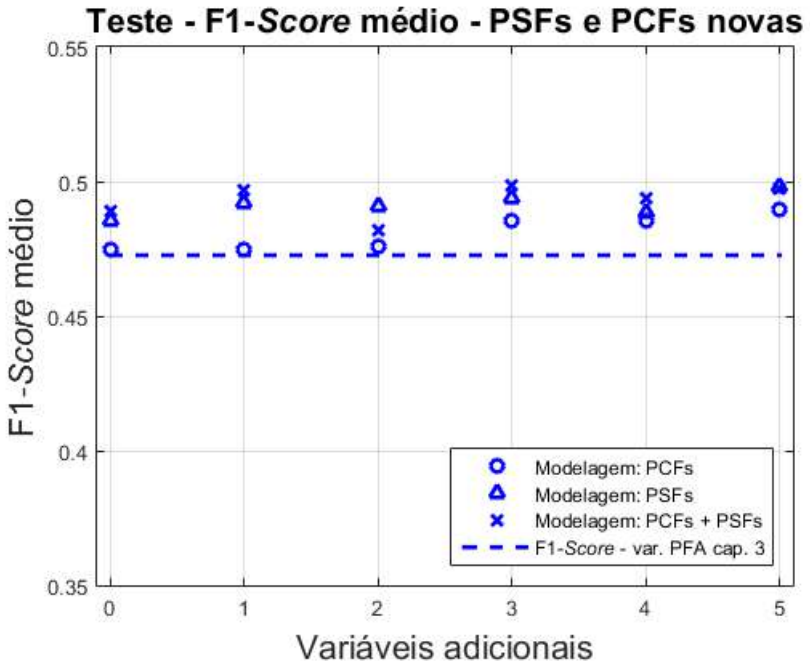


Figura 5.13: Com a seleção de variáveis adicionando as novas PCFs e PSFs, observa-se que todos os blocos de conjuntos de variáveis alcançaram resultados melhores que o Experimento 7 com as variáveis selecionadas na seção 3.4.3. O F1-*Score* médio do conjunto de melhor desempenho é $0,4987$.

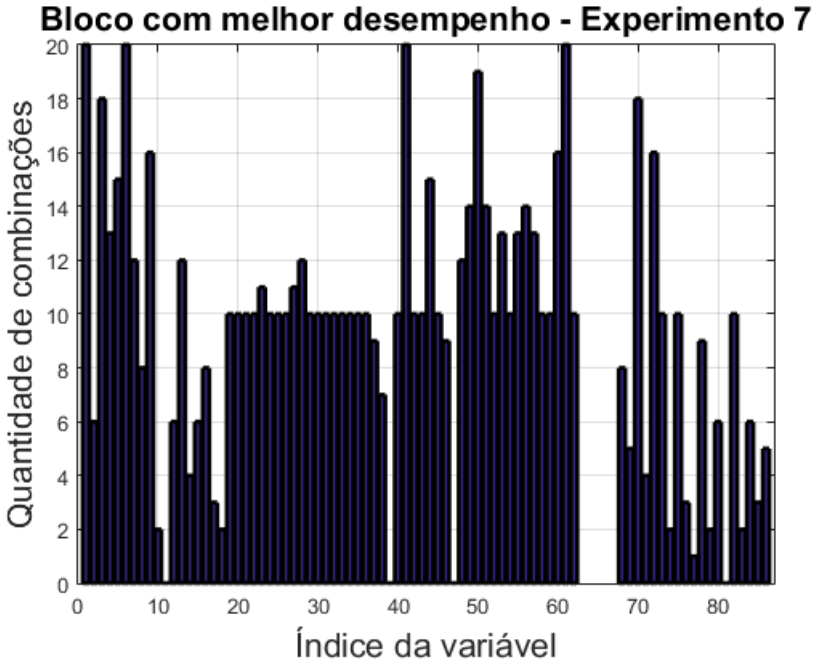


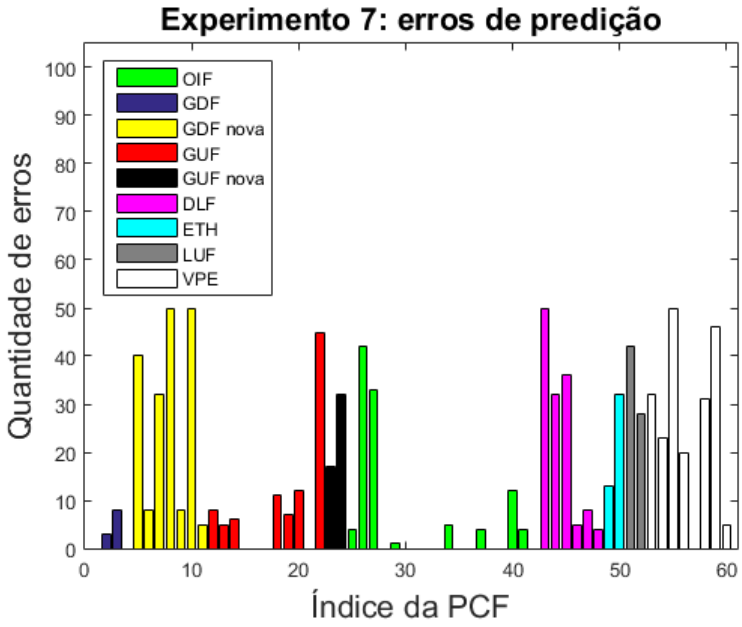
Figura 5.14: No bloco de conjuntos de variáveis com melhor desempenho médio, observa-se que 25 variáveis, listadas na Tabela 5.15, estão presentes em mais de 10 combinações de grupos. Há, também, 4 variáveis, listadas na Tabela 5.14, não selecionadas em nenhuma das combinações.

Tabela 5.15: PFA, Experimento 7: variáveis presentes em mais de 50% das combinações.

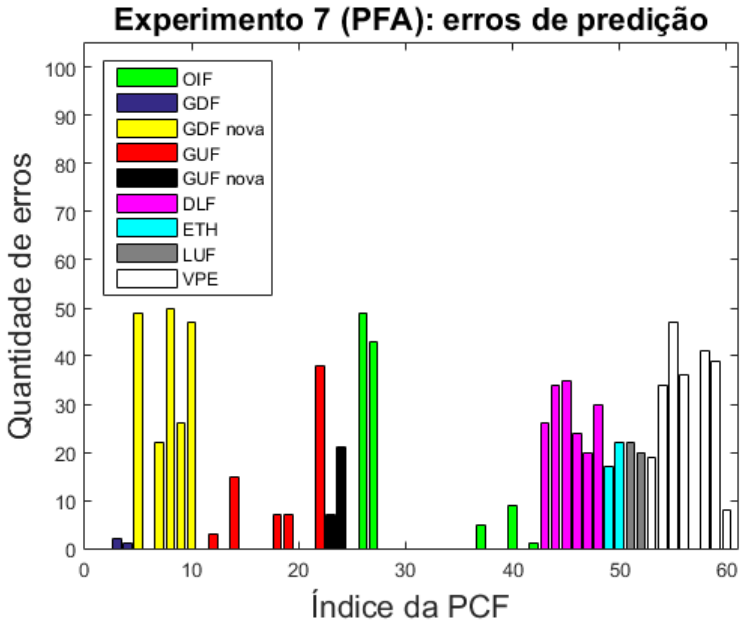
Índice	**Identificação**	**Descrição**	**Combinações**
1	FIT-001	Vazão de gás combustível	20
3	TI-002	Temperatura do *header* de óleo mineral	18
4	TIT-001	Saída de água quente do WHRU	13
5	TIT-002	Temperatura do WHRU	15
6	TT-001	Temperatura ambiente	20
7	TT-002	Temperatura do gás combustível	12
9	VE-001	Vibração na entrada do GG	16
13	VE-005	Vib. PT disc. end. Y	12
23	TI-003-04	Termopar 04 do perfil da exaustão	11
27	TI-003-08	Termopar 08 do perfil da exaustão	11
28	TI-003-09	Termopar 09 do perfil da exaustão	12
41	PDT-002	Pressão diferencial - Filtro de gás combustível	20
44	PIT-001	Pressão	15
48	PT-003	Gás Comb.	12
49	ST-001	Rot. Com. / Turb. Baixa	14
50	ST-002	Rotação Motor de Arranque	19
51	ST-003	Rot. Com. / Turb. Alta	14
53	TE-001	Exaustão GG	13
55	ET-001	Voltagem ab	13
56	ET-001	Voltagem bc	14
57	ET-001	Voltagem ca	13
60	JY-001	Energia Real	16
61	JQ-001	*Power Factor*	20
70	TI-008	Exaustão GG	18
72	TI-010	PT Temperatura Mancal de Escora	16

Tabela 5.16: PFA, Experimento 7: variáveis não selecionadas nas combinações.

Índice	**Identificação**	**Descrição**
11	VE-003	Vibração na turbina do GG
39	TI-006	Gerador temp. enrolamento L2-1
47	PT-003	Entrada WHRU
81	TI-019	Gerador ar de resfriamento (frio) NDE



(a)



(b)

Figura 5.15: Erros de predição em casos de PCF no Experimento 7: (a) observa-se que o perfil de identificação dos casos de GDF, GUF e OIF é melhor que o mostrado na Figura 3.5(b), referente ao experimento realizado na seção 3.4.3 com 81 variáveis. O conjunto de novas PCFs apresentou muitos erros de predição; (b) em comparação com (a), nota-se que, após a aplicação do PFA, os classificadores treinados foram capazes de melhorar seu desempenho em relação ao Experimento 7 com as variáveis selecionadas na seção 3.4.3, com exceção dos novos casos de GDF e dos casos de VPE. Esta capacidade resultou num desempenho global melhor, conforme observado na Figura 5.13.

5.4.6 Análise conjunta dos novos experimentos

Com a aplicação do método de seleção de variáveis utilizando PFA nos novos experimentos, verificou-se que a seleção de variáveis resulta numa distribuição dos erros de classificação mais uniforme entre os casos de PCF. Este resultado mostra que este método proporciona uma melhor caracterização dos casos analisados.

Analogamente aos resultados dos novos experimentos sem utilizar o PFA, os resultados do Experimento 5, a comparação entre os Experimento 3 e 6 e a comparação do Experimento 4 com o Experimento 7 indicam que a adição dos novos casos de PSF proporcionou uma melhoria na identificação das PCFs. A inserção de novos casos de PSF também resultou no aumento do número de variáveis selecionadas para representação, aumentando o custo computacional para o projeto dos classificadores.

5.5 Análise dos casos de PSF

Conforme já indicado em [3], existem casos indicados como PSF que se assemelham aos casos de PCF. Estas ocorrências podem ser decorrentes, por exemplo, de padrões de operação degradados que não causaram falhas iminentes. Estas PSFs foram identificadas através da análise dos casos treinados que foram categorizados incorretamente pelos classificadores resultantes [3].

Com o aumento da quantidade de PSFs e da diversidade dos tipos de PCF estudadas, verificou-se a necessidade de identificar quais casos de PSF são significativos considerando o critério adotado em [3]. Desta forma, os Experimentos 6 e 7 foram complementados para conter esta avaliação. Seguem as etapas que constituem este complemento:

1. Ampliação do conjunto de PSFs analisadas, incluindo todos os 193 casos referentes ao período de entre 2010 e 2012;
2. Treinamento de classificadores utilizando todas as 81 variáveis disponíveis;
3. Identificação de casos de PSF classificados incorretamente na fase de treino;
4. Seleção do subconjunto de PSFs para realização dos experimentos;
5. Treinamento de classificadores utilizando subconjunto de PSFs selecionado e todas as 81 variáveis disponíveis;
6. Aplicação do algoritmo de seleção de variáveis do capítulo 3;
7. Comparação dos resultados com os Experimentos 6 e 7 originais.

As seções 5.5.1 e 5.5.2 apresentam os resultados da análise dos casos de PSF nos Experimentos 6 e 7, respectivamente.

5.5.1 Experimento 6 com análise de PSFs

Utilizando as 81 variáveis disponíveis e todos os casos de PSF, foram executadas as 5 repetições do treinamento de classificadores das 20 combinações dos seis grupos de PCFs e PSFs, seguindo metodologia da seção 2.4.3. Após o treinamento dos classificadores utilizando os 523 casos de PSF disponíveis, a quantidade de erros durante a fase de treinamento foi adquirida e apresentada na Figura 5.16. Observa-se que há maior concentração de casos sem erros de classificação, correspondendo a 362 PSFs. Desta forma, estes casos de PSF foram selecionados.

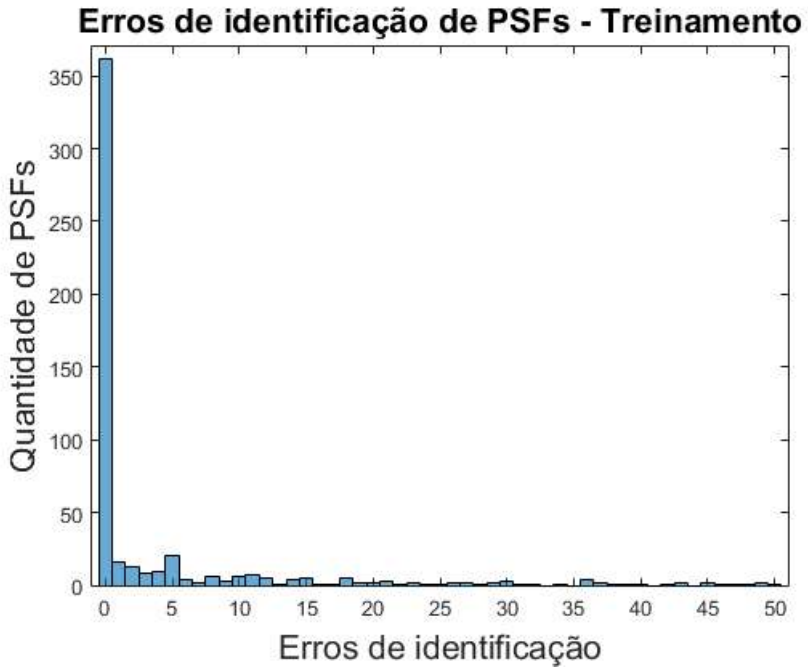


Figura 5.16: Erros de classificação de PSFs durante fase de treinamento no Experimento 6: observa-se que há maior concentração de casos sem erros de classificação, correspondendo a 362 PSFs.

Com este novo subconjunto de PSFs, seguiu-se com o treinamento de novos classificadores utilizando todas as 81 variáveis disponíveis. O F1-*Score* médio destes classificadores é 0,5112. Este valor corresponde a uma taxa de acerto médio de 86,57% para as PSFs e de 74,29% para as PCFs.

Por fim, aplicando o PFA, o Experimento 6 foi realizado novamente com todos os 18 blocos de conjuntos de variáveis obtidos pela combinação das 3 modelagens estatísticas propostas. A distribuição dos casos de PCF manteve-se conforme a Tabela 5.5. Com a ordenação temporal do novo subconjunto de PSFs, a distribuição dos casos de PSF é mostrada na Tabela 5.17.

Tabela 5.17: Distribuição das 362 PSFs selecionadas.

	G1	G2	G3	G4	G5	G6
TGA-PSF	12	12	11	12	11	11
TGB-PSF	16	16	17	16	16	17
TGC-PSF	17	17	17	17	17	16
TGD-PSF	16	16	15	15	16	16

A comparação de desempenho dos blocos de classificadores é realizada através do valor da média do F1-*Score* obtido nas 5 execuções das 20 combinações dos 6 grupos de PSFs e PCFs.

Na Figura 5.17, observa-se que 2 blocos de conjuntos de variáveis alcançaram resultados melhores que o Experimento 6 original utilizando as variáveis selecionadas na seção 5.4.4. O bloco de melhor desempenho é composto pelos conjuntos de variáveis, com 4 variáveis adicionais, selecionados através da modelagem estatística a partir de casos de PSFs e PCFs em conjunto. A quantidade total de variáveis destes conjuntos variou entre 18 e 21. O valor do F1-*Score* médio deste conjunto é 0,5218. Este valor corresponde a uma taxa de acerto médio de 87,06% para as PSFs e de 74,67% para as PCFs.

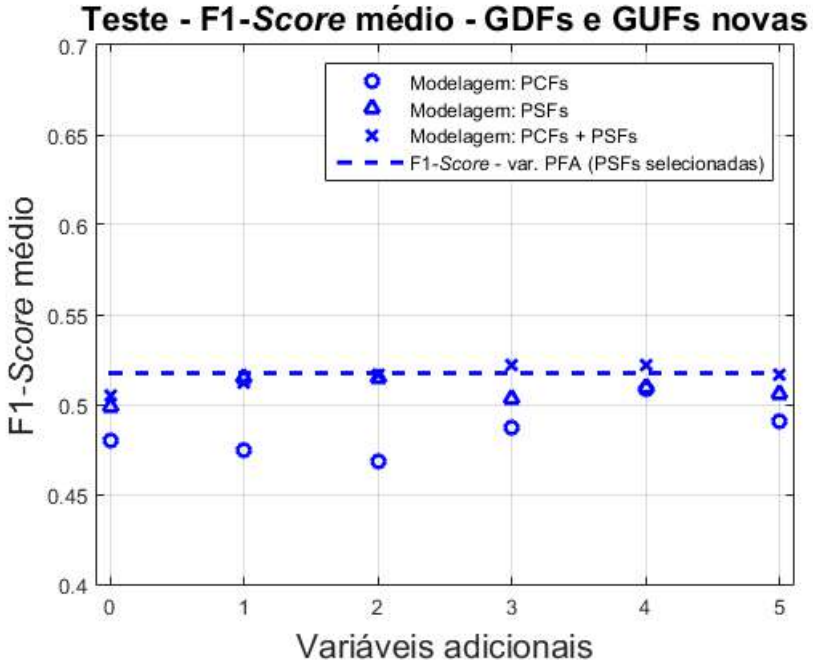


Figura 5.17: Com a seleção de variáveis adicionando as novas GDFs, GUFs e utilizando o grupo de PSFs selecionado, observa-se que 2 blocos de conjuntos de variáveis alcançaram resultados melhores que o Experimento 6 com as variáveis selecionadas na seção 5.4.4. O F1-*Score* médio do conjunto de melhor desempenho é $0,5218$.

Verificou-se que as variáveis selecionadas variaram bastante para cada combinação de grupos. A quantidade de combinações em que cada uma das variáveis foi

selecionada é apresentada na Figura 5.18. A Tabela 5.18 lista as 19 variáveis presentes em mais de 50% das combinações de grupos. Neste bloco, 29 variáveis, listadas na Tabela 5.19, não foram selecionadas em nenhuma das combinações.

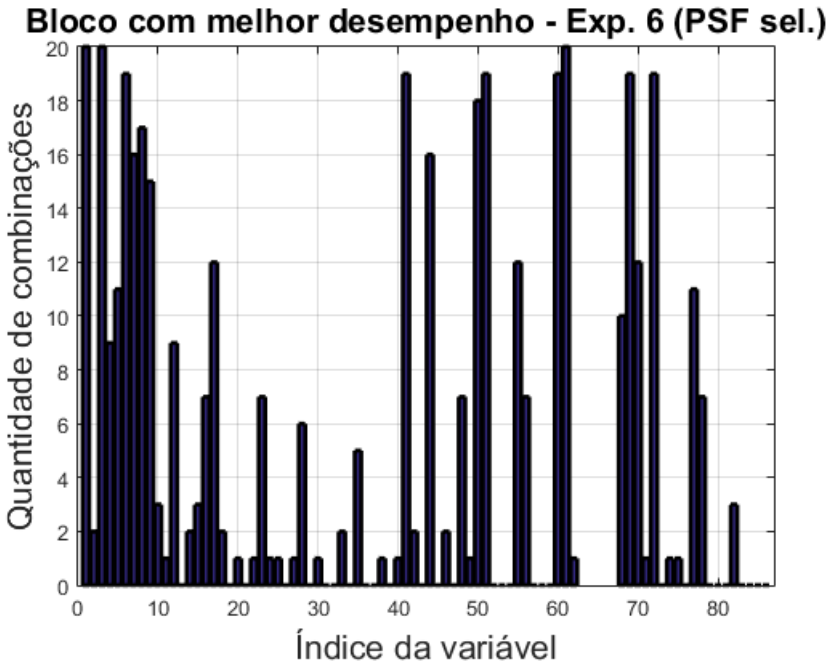


Figura 5.18: No bloco de conjuntos de variáveis com melhor desempenho médio, observa-se que 19 variáveis, listadas na Tabela 5.18, estão presentes em mais de 10 combinações de grupos. Há, também, 29 variáveis, listadas na Tabela 5.19, não selecionadas em nenhuma das combinações.

Tabela 5.18: PFA, Experimento 6 com seleção de PSFs: variáveis presentes em mais de 50% das combinações.

Índice	**Identificação**	**Descrição**	**Combinações**
1	FIT-001	Vazão de gás combustível	20
3	TI-002	Temperatura do *header* de óleo mineral	20
5	TIT-002	Temperatura do WHRU	11
6	TT-001	Temperatura ambiente	19
7	TT-002	Temperatura do gás combustível	16
8	TT-003 A/B/C/D	Temperatura do *manifold* de gás combustível	17
9	VE-001	Vibração na entrada do GG	15
17	ZE-002 A/B/C/D	Vib. PT AXIAL	12
41	PDT-002	Pressão diferencial - Filtro de gás combustível	19
44	PIT-001	Pressão	16
50	ST-002	Rotação Motor de Arranque	18
51	ST-003	Rot. Com. / Turb. Alta	19
55	ET-001	Voltagem ab	12
60	JY-001	Energia Real	19
61	JQ-001	*Power Factor*	20
69	TI-007 A/B/C/D	Gerador temp. enrolamento L1-3	19
70	TI-008	Exaustão GG	12
72	TI-010	PT Temperatura Mancal de Escora	19
77	TI-015 A/B/C/D	*Gearbox* HSS Temperatura mancal NDE	11

Tabela 5.19: PFA, Experimento 6 com seleção de PSFs: variáveis não selecionadas nas combinações.

Índice	Identificação	Descrição
13	VE-005 A/B/C/D	Vib. PT disc. end. Y
19	ZE-004 A/B/C/D	*Gearbox* HSS Axial
21	TI-003-02	Termopar 02 do perfil da exaustão
26	TI-003-07	Termopar 07 do perfil da exaustão
29	TI-003-10	Termopar 10 do perfil da exaustão
31	TI-003-12	Termopar 12 do perfil da exaustão
32	TI-003-13	Termopar 13 do perfil da exaustão
34	TI-003-15	Termopar 15 do perfil da exaustão
36	TI-003-17	Termopar 17 do perfil da exaustão
37	TI-004 A/B/C/D	Gerador temp. enrolamento L2-3
39	TI-006 A/B/C/D	Gerador temp. enrolamento L2-1
43	PI-001 A/B/C/D	Pressão header óleo mineral
45	PT-001 A/B/C/D	Pressão P1
47	PT-003 A/B/C/D	Entrada WHRU
52	ST-004 A/B/C/D	Rotação PT
53	TE-001 A/B/C/D	Exaustão GG
54	IT-001 A/B/C/D	Corrente de excitação do campo
57	ET-001 A/B/C/D	Voltagem ca
58	JT-001 A/B/C/D	Potência Reativa
59	JT-002 A/B/C/D	*Active Power*
73	TI-011 A/B/C/D	PT Temperatura mancal DE
76	TI-014 A/B/C/D	*Gearbox* LSS Temperatura mancal NDE
79	TI-017 A/B/C/D	Gerador ar de resfriamento (quente) DE
80	TI-018 A/B/C/D	Gerador ar de resfriamento (quente) NDE
81	TI-019 A/B/C/D	Gerador ar de resfriamento (frio) NDE
83	TI-021 A/B/C/D	Gerador temp. enrolamento L1-1
84	TI-022 A/B/C/D	Gerador temp. enrolamento L1-2
85	TI-023 A/B/C/D	Gerador temp. enrolamento L3-1
86	TI-024 A/B/C/D	Gerador temp. enrolamento L3-2

5.5.2 Experimento 7 com análise de PSFs

Utilizando as 81 variáveis disponíveis e todos os casos de PSF, foram executadas as 5 repetições do treinamento de classificadores das 20 combinações dos seis grupos de PCFs e PSFs, seguindo metodologia da seção 2.4.3. Após o treinamento dos classificadores utilizando os 523 casos de PSF disponíveis, a quantidade de erros durante a fase de treinamento foi adquirida e apresentada na Figura 5.19. Observa-se que há maior concentração de casos sem erros de classificação, correspondendo a 363 PSFs. Desta forma, estes casos de PSF foram selecionados.

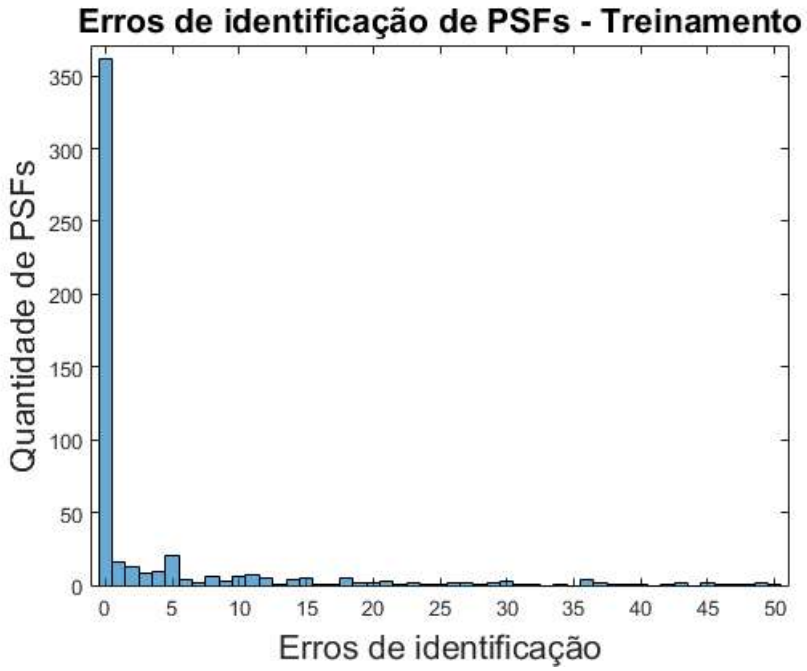


Figura 5.19: Erros de classificação de PSFs durante fase de treinamento no Experimento 7: observa-se que há maior concentração de casos sem erros de classificação, correspondendo a 363 PSFs.

Com este novo subconjunto de PSFs, seguiu-se com o treinamento de novos classificadores utilizando todas as 81 variáveis disponíveis. O F1-*Score* médio destes classificadores é 0,5176. Este valor corresponde a uma taxa de acerto médio de 82,07% para as PSFs e de 72,83% para as PCFs.

Por fim, aplicando o PFA, o Experimento 7 foi realizado novamente com todos os 18 blocos de conjuntos de variáveis obtidos pela combinação das 3 modelagens estatísticas propostas. A distribuição dos casos de PCF manteve-se conforme a Tabela 5.8. Com a ordenação temporal do novo subconjunto de PSFs, a distribuição dos casos de PSF é mostrada na Tabela 5.20.

Tabela 5.20: Distribuição das 363 PSFs selecionadas.

	G1	**G2**	**G3**	**G4**	**G5**	**G6**
TGA-PSF	13	13	12	13	12	12
TGB-PSF	14	14	15	14	15	15
TGC-PSF	16	16	16	16	16	16
TGD-PSF	18	18	17	18	17	17

A comparação de desempenho dos blocos de classificadores é realizada através do valor da média do F1-*Score* obtido nas 5 execuções das 20 combinações dos 6 grupos de PSFs e PCFs.

Na Figura 5.20, observa-se que 16 blocos de conjuntos de variáveis alcançaram resultados melhores que o Experimento 7 original utilizando as variáveis selecionadas

na seção 5.4.5. O bloco de melhor desempenho é composto pelos conjuntos de variáveis, com 5 variáveis adicionais, selecionados através da modelagem estatística a partir de casos de PSFs e PCFs em conjunto. A quantidade total de variáveis destes conjuntos variou entre 20 e 23. O valor do F1-*Score* médio deste conjunto é 0,5256. Este valor corresponde a uma taxa de acerto médio de 82,72% para as PSFs e de 72,87% para as PCFs.

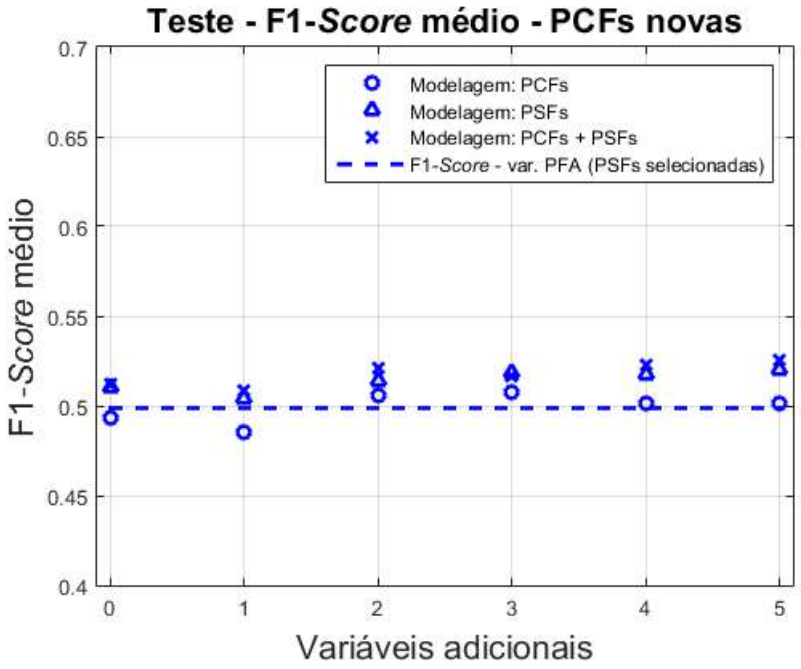


Figura 5.20: Com a seleção de variáveis adicionando as novas PCFs e utilizando o grupo de PSFs selecionado, observa-se que 16 blocos de conjuntos de variáveis alcançaram resultados melhores que o Experimento 7 com as variáveis selecionadas na seção 5.4.5. O F1-*Score* médio do conjunto de melhor desempenho é 0,5256.

Verificou-se que as variáveis selecionadas variaram bastante para cada combinação de grupos. A quantidade de combinações em que cada uma das variáveis foi selecionada é apresentada na Figura 5.21. A Tabela 5.21 lista as 19 variáveis presentes em mais de 50% das combinações de grupos. Neste bloco, 30 variáveis, listadas na Tabela 5.22, não foram selecionadas em nenhuma das combinações.

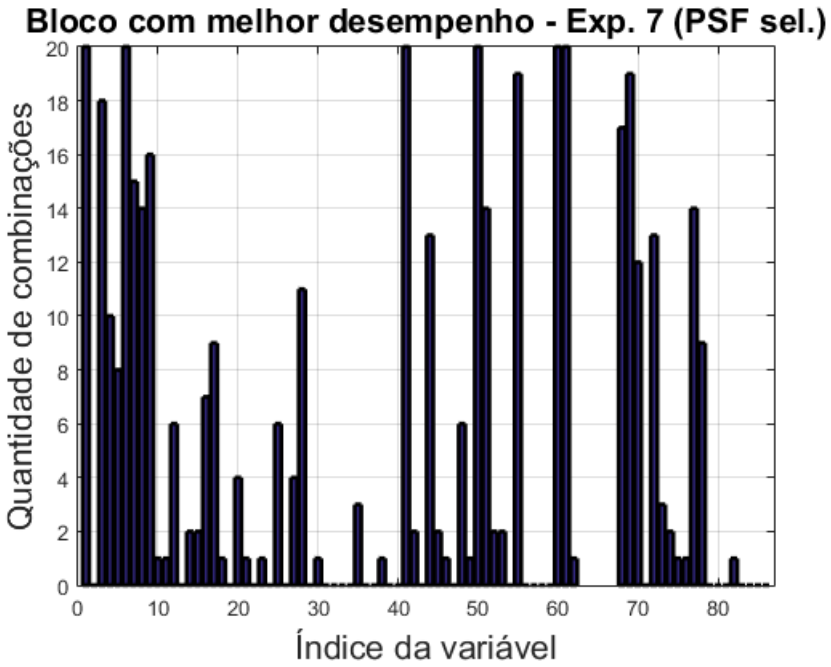


Figura 5.21: No bloco de conjuntos de variáveis com melhor desempenho médio, observa-se que 19 variáveis, listadas na Tabela 5.21, estão presentes em mais de 10 combinações de grupos. Há, também, 30 variáveis, listadas na Tabela 5.22, não selecionadas em nenhuma das combinações.

Tabela 5.21: PFA, Experimento 7 com seleção de PSFs: variáveis presentes em mais de 50% das combinações.

Índice	**Identificação**	**Descrição**	**Combinações**
1	FIT-001	Vazão de gás combustível	20
3	TI-002	Temperatura do *header* de óleo mineral	18
6	TT-001	Temperatura ambiente	20
7	TT-002	Temperatura do gás combustível	15
8	TT-003 A/B/C/D	Temperatura do *manifold* de gás combustível	14
9	VE-001	Vibração na entrada do GG	16
28	TI-003-09	Termopar 09 do perfil da exaustão	11
41	PDT-002	Pressão diferencial - Filtro de gás combustível	20
43	PI-001 A/B/C/D	Pressão header óleo mineral	13
50	ST-002	Rotação Motor de Arranque	20
51	ST-003	Rot. Com. / Turb. Alta	14
55	ET-001	Voltagem ab	19
60	JY-001	Energia Real	20
61	JQ-001	*Power Factor*	20
68	IT-002 A/B/C/D	Corrente	17
69	TI-007 A/B/C/D	Gerador temp. enrolamento L1-3	19
70	TI-008	Exaustão GG	12
72	TI-010	PT Temperatura Mancal de Escora	13
77	TI-015 A/B/C/D	*Gearbox* HSS Temperatura mancal NDE	14

Tabela 5.22: PFA, Experimento 7 com seleção de PSFs: variáveis não selecionadas nas combinações.

Índice	Identificação	Descrição
2	TI-001 A/B/C/D	Temperatura do tanque de óleo sintético
13	VE-005 A/B/C/D	Vib. PT disc. end. Y
19	ZE-004 A/B/C/D	*Gearbox* HSS Axial
22	TI-003-03	Termopar 03 do perfil da exaustão
24	TI-003-05	Termopar 05 do perfil da exaustão
26	TI-003-07	Termopar 07 do perfil da exaustão
29	TI-003-10	Termopar 10 do perfil da exaustão
31	TI-003-12	Termopar 12 do perfil da exaustão
32	TI-003-13	Termopar 13 do perfil da exaustão
33	TI-003-14	Termopar 14 do perfil da exaustão
34	TI-003-15	Termopar 15 do perfil da exaustão
36	TI-003-17	Termopar 17 do perfil da exaustão
37	TI-004 A/B/C/D	Gerador temp. enrolamento L2-3
39	TI-006 A/B/C/D	Gerador temp. enrolamento L2-1
40	PDT-001 A/B/C/D	Pressão diferencial - Entrada GG
43	PI-001 A/B/C/D	Pressão *header* óleo mineral
47	PT-003 A/B/C/D	Entrada WHRU
54	IT-001 A/B/C/D	Corrente de excitação do campo
56	ET-001 A/B/C/D	Voltagem bc
57	ET-001 A/B/C/D	Voltagem ca
58	JT-001 A/B/C/D	Potência Reativa
59	JT-002 A/B/C/D	*Active Power*
71	TI-009 A/B/C/D	PT Temperatura Mancal NDE
79	TI-017 A/B/C/D	Gerador ar de resfriamento (quente) DE
80	TI-018 A/B/C/D	Gerador ar de resfriamento (quente) NDE
81	TI-019 A/B/C/D	Gerador ar de resfriamento (frio) NDE
83	TI-021 A/B/C/D	Gerador temp. enrolamento L1-1
84	TI-022 A/B/C/D	Gerador temp. enrolamento L1-2
85	TI-023 A/B/C/D	Gerador temp. enrolamento L3-1
86	TI-024 A/B/C/D	Gerador temp. enrolamento L3-2

5.5.3 Análise conjunta dos novos experimentos

Com a aplicação do método de seleção de PSFs nos novos Experimentos 6 e 7, verificou-se que o F1-*Score* obtido foi superior aos obtidos na seção 5.4. Este resultado mostra, novamente, que a metodologia apresentada em [3] é válida para filtrar casos PSF que possuam comportamentos similares às PCFs.

5.5.4 Comparação com o Experimento 2

A comparação direta entre os Experimentos 3 a 7 com o Experimento 2 através do F1-*Score* não é adequada, pois esta métrica, conforme verificado nas equações (3.9),(3.10) e (3.11), é dependente diretamente da relação entre a quantidade total de PCFs e de PSFs. Como o total de PSFs aumentou numa proporção maior que o total de PCFs, os F1-*Score* destes experimentos tendem a ser menores que o obtido no Experimento 2 mesmo quando alcançam percentuais de acerto médio superiores.

Dentre os novos experimentos apresentados neste capítulo, os Experimentos 6 e 7 representam o maior aumento de complexidade de classificação devido ao grande número de novas PSFs e à adição de novos tipos de PCF. A comparação de resultados destes experimentos com o Experimento 2 permitiria validar a robustez da metodologia estudada nesta dissertação em comparação com os resultados de [3], considerando o incremento de complexidade imposto pela expansão do banco de dados original.

Para viabilizar esta comparação, utilizando os classificadores de melhor desempenho das seções 5.5.1 e 5.5.2, as PSFs dos conjuntos de testes foram subamostradas para manter a relação entre casos de PSF e PCF iguais à proporção do Experimento 2. Para garantir que o processo de subamostragem represente a diversidade de casos de PSF presentes nas 20 combinações dos 3 grupos de PSFs, cada experimento deve ser repetido com subgrupos de PSFs selecionados de forma aleatória. Considerando que cada combinação i possua quantidade de PCFs igual a $numPCF_i$, a quantidade de PSFs selecionadas $numPSFsel_i$ é calculada por

$$numPSFsel_i = \left\lceil \frac{numPCF_i \times 70}{33} \right\rceil, \tag{5.1}$$

onde $\lceil \frac{a}{b} \rceil$ retorna o primeiro inteiro acima do quociente de a dividido por b. O resultado final do experimento é calculado através da média dos F1-*Score* obtidos em cada uma das repetições realizadas para cada uma das combinações.

Realizando 100 repetições desta subamostragem aleatória, o F1-*Score* médio do Experimento 6 calculado é igual a 0,7249. Este valor corresponde a uma taxa de acerto médio de 85,3% para as PSFs e de 74,67% para as PCFs. Já o F1-*Score* médio do Experimento 7 calculado é valor igual a 0,6908. Este valor corresponde a uma taxa de acerto médio de 82,44% para as PSFs e de 72,87% para as PCFs.

Os resultados obtidos indicam que a metodologia aplicada mantém um desempenho bastante próximo aos resultados do Experimento 2 obtidos no Capítulo 2, com F1-*Score* igual a 0,7212, e no Capítulo 3, com F1-*Score* igual a 0,7425, mesmo após a expansão do banco de dados.

5.6 Conclusão

Este capítulo detalhou os novos casos de PCF e PSF obtidos após a aquisição de novos dados de operação e eventos de manutenção dos turbogeradores, referentes ao período de janeiro de 2013 a julho de 2016. A ampliação dos bancos de dados é de grande importância para esta pesquisa, pois permite analisar de forma mais abrangente a operação dos turbogeradores e avaliar os resultados obtidos utilizando o banco de dados original.

Um total de 27 PCFs e 330 PSFs foram identificadas neste novo banco de dados. Além do aumento considerável da quantidade de PSFs, identificou-se que a distribuição dos casos de PCF neste novo período difere bastante da distribuição no período entre 2010 e 2012. Além de novos casos de GDF e GUF, o novo período de operação contou com tipos diferentes de PCF: DLF, ETH, LUF e VPE.

Após o devido tratamento, conforme seção 2.4, as novas PSFs e PCFs foram utilizadas para testar, na seção 5.3, os classificadores treinados com os dados refentes ao período de 2010 a 2012. Como os resultados não foram satisfatórios, novos classificadores foram treinados, nas seções 5.3.1 e 5.4, tanto com as variáveis selecionadas em 3.4.3 como com as variáveis selecionadas por uma nova aplicação do PFA, utilizando o banco de dados completo.

Analisando os resultados obtidos, conclui-se que a aplicação do PFA, em geral, proporcionou melhora na identificação dos classificadores. A distribuição das taxas de erro fica mais uniforme e com valor médio reduzido.

A utilização dos novos casos de PSF proporcionou uma melhora na taxa de identificação das PCFs. No entanto, percebe-se que alguns casos de PCF ainda possuem taxas de erro muito altas. Este resultado pode ser consequência da existência de casos de PSF que sejam semelhantes a estes casos de PCF, da mesma forma que foi apresentado em [3]. Aplicou-se, então, nos Experimentos 6 e 7 a metodologia de seleção de casos de PSF descrita em [3]. Ao utilizar os novos conjuntos de PSFs selecionados, os resultados foram superiores aos experimentos da seção 5.4, que utilizam todos os casos de PSF disponíveis. Reforça-se, assim, a necessidade de aprofundar o estudo da efetividade da classificação dos casos de operação considerados normais.

Por fim, através da subamostragem dos casos de PSF dos conjuntos de teste, geraram-se resultados dos Experimentos 6 e 7 com proporções entre casos de PCF e PSF equivalentes à proporção do Experimento 2. Verificou-se que o desempenho destes experimentos com o banco de dados estendido manteve-se bastante próximo ao desempenho do Experimento 2 dos Capítulos 2 e 3. Esta manutenção de desempenho, frente ao incremento de complexidade imposto pelo aumento do banco de dados, constitui um resultado relevante desta dissertação. Ela indica que os métodos propostos são robustos e capazes de lidar com a variação e diversificação

operacionais dos equipamentos estudados.

Capítulo 6

Conclusão e Trabalhos Futuros

Esta dissertação estudou a detecção de falhas em partidas de turbogeradores. Conforme apresentado no Capítulo 2 e proposto em [3], classificadores são desenvolvidos para a identificação de PCFs e PSFs. Adotou-se a premissa de que os dados de processo de 24 horas de operação contínuas que antecedem as paradas observadas podem indicar antecipadamente o estado do equipamento. A reprodução destes experimentos permitiu a familiarização com as técnicas utilizadas e o entendimento dos bancos de dados utilizados. Os resultados iniciais obtidos sugerem que a premissa proposta é razoável. No entanto, percebeu-se que o banco de dados de variáveis de processo não havia sido explorado de forma plena. Com isso, o foco principal deste trabalho foi o tratamento e a análise detalhada dos bancos de dados buscando melhorias na separação destas classes.

No Capítulo 3, objetivou-se determinar o conjunto de variáveis ideal para representação dos casos de operação analisados. Realizando aproximações estatísticas a partir dos conjuntos de amostras, o PCA foi aplicado para analisar o potencial de redução dimensional determinando o número de CPs que representam as PCFs e PSF estudadas mantendo retenção de 90% da energia. Esta análise considerou tanto o conjunto de 22 variáveis utilizado no Capítulo 2 como o conjunto completo de 81 variáveis disponibilizado.

Para seleção de variáveis em si, implementou-se a técnica PFA nos dois conjuntos de variáveis analisados. Notou-se que a quantidade de variáveis selecionadas, para cada subconjunto de casos de modelagem, manteve-se próxima à quantidade de CPs determinada pela aplicação do PCA. Além disso, destaca-se que a utilização desta técnica permitiu a seleção de variáveis de forma semisupervisionada com resultados equivalentes, partindo do conjunto de 22 variáveis, e melhores que os resultados apresentados no Capítulo 2 partindo do conjunto de 81 variáveis. Este resultado é bastante importante, considerando que nem sempre é possível determinar a relação direta entre padrões de operação com as variáveis monitoradas. E, mesmo quando esta relação é conhecida, ainda é possível que haja variáveis de processo que resultem

em melhorias na tarefa de classificação.

A busca por uma representação mais eficiente dos casos de operação foi estendida no Capítulo 4. O KPCA possibilita, através da escolha de diferentes funções de *kernel*, o mapeamento de relações não lineares no conjunto de variáveis original. A projeção destes dados nas direções das componentes principais destes domínios transformados tem potencial de gerar características que permitam uma separação mais eficiente das PCFs e PSFs. Ao aplicar esta técnica, com funções de *kernel* polinomial e gaussiano, nos dados de processo dos turbogeradores e gerar classificadores conforme a metodologia do Capítulo 2, verificou-se que os novos classificadores resultantes obtiveram resultados piores que o uso direto das séries temporais. Deve-se ressaltar que o número máximo de características geradas pelo KPCA é diretamente influenciado pela quantidade de casos de PCF e PSF disponíveis. Com isso, os classificadores foram treinados com um conjunto de pontos bastante reduzido, que pode ter influenciado negativamente os resultados obtidos.

O Capítulo 5, inicialmente, detalhou a expansão dos bancos de dados dos turbogeradores. Os dados referentes ao período de janeiro de 2013 a julho de 2016 foram adquiridos e tratados seguindo todos os passos descritos no Capítulo 2. Após análise do banco de dados de eventos de manutenção, percebeu-se que o perfil dos casos de PCF modificou-se em relação ao período de janeiro de 2010 a 2012, com ocorrência de outros tipos de falha e modificação na proporção dos tipos de falha já anotados. Além disso, verificou-se que a proporção de PSFs em relação às PCFs cresceu significativamente.

Os novos casos de PCF e PSF foram utilizados como conjunto de testes de classificadores parametrizados e treinados a partir dos dados provenientes do período de 2010 a 2012. Embora a identificação de PSFs de forma isolada tenha mantido desempenho compatível com os resultados dos Capítulos 2 e 3, os classificadores não foram capazes de separar adequadamente os novos casos de PCF de casos de PSF, sugerindo que a dinâmica dos novos casos de PCF seja diferenciada dos casos anteriores.

Para avaliar o peso da inserção dos novos casos de operação no conjunto inicial, novos classificadores foram desenvolvidos. Numa primeira etapa, os conjuntos de variáveis selecionados no Capítulo 3 foram utilizados. Na sequência, o PFA foi aplicado novamente para seleção de variáveis. O método de seleção de variáveis mostrou consistentemente que ele é capaz de distribuir de forma mais uniforme os erros de predição dentre os tipos de PCF estudados. No entanto, percebe-se que alguns tipos de PCF não foram separados adequadamente das PSFs, acarretando resultados globais piores em alguns casos.

Com a inserção dos novos de casos de PSF, em quantidade maior que o triplo ao número de casos tratados inicialmente, o número de erros de identificação dos casos

de PCF foi reduzido, com exceção de alguns casos novos de PCF. Houve também, para certas combinações de grupos de casos de operação, um aumento significativo de variáveis selecionadas. Além de possuir implicação direta no tempo necessário para treinamento dos novos classificadores, este aumento aponta que a variedade de novos casos de PSF demanda um número maior de CPs necessárias para sua representação, mantendo 90% da energia retida.

Com o intuito de excluir da análise os casos de PSF que possuem comportamento semelhante ao das PCFs, aplicou-se a metodologia de exclusão de PSFs apresentada em [3] nos experimentos que utilizam o banco de dados expandido. Esta seleção resultou numa melhoria de performance equivalente à melhoria mostrada no Capítulo 2 e em [3].

Tendo em vista que a definição dos casos de PSF e PCF são dependentes de uma base de dados anotada, os resultados desta dissertação, em conjunto com os casos de PSF descartados em [3], indicam que os critérios para definição de PSFs podem ser revisados e aprimorados. Desta forma, o diagnóstico dos modos de funcionamento dos turbogeradores, com a metodologia apresentada nesta dissertação, pode ser refinado.

Outro resultado relevante desta pesquisa consiste na manutenção de performance dos experimentos desenvolvidos utilizando o banco de dados estendido em comparação com os resultados dos Capítulos 2 e 3. Mantendo-se a proporção de casos de PCF e de PSF equivalente à proporção utilizada no Capítulo 3, verificou-se que os resultados obtidos são bastante próximos. Com isso, conclui-se que as metodologias propostas em [3] e aprofundadas nesta dissertação são capazes de lidar com a diversidade operacional dos turbogeradores estudados.

Este trabalho não esgota as possibilidades de melhoria para o problema apresentado. O estudo dos pontos indicados a seguir pode enriquecer a pesquisa realizada:

- Avaliar a variação de desempenho das técnicas de PCA estudadas através da utilização de valores diversificados de variabilidade retida;
- Aplicação de novas funções de *kernel* [25, 28] e desenvolvimento de métodos para definição de seus parâmetros [29];
- Utilização de outros tipos de classificadores e o tratamento das séries de dados para viabilizar sua utilização;
- Considerando o grande aumento de casos de PSFs, abordar estes casos através de um classificador de classe única [30];
- Aplicação de segmentação e análise localizada de séries temporais, buscando regras de funcionamento [31].

Referências Bibliográficas

[1] HENG, A., ZHANG, S., TAN, A. C., et al. "Rotating machinery prognostics: State of the art, challenges and opportunities", *Mechanical systems and signal processing*, v. 23, n. 3, pp. 724-739, Abr. 2009.

[2] MACHADO, M. M., MANGUINHO, D. A. P. M., VALLAND, A., et al. "RUL modeling for turbo generators of a FPSO: Alternatives and challenges". In: *Proceedings of the RIO OIL & GAS 2014*, Rio de Janeiro, Set. 2014.

[3] SANTOS, I., MACHADO, M., RUSSO, E., et al. "Big Data Analytics for Predictive Maintenance Modeling: Challenges and Opportunities". In: *Proceedings of the Offshore Technology Conference 2015*, Rio de Janeiro, Out. 2015.

[4] MARWALA, T. *Condition monitoring using computational intelligence methods: applications in mechanical and electrical systems*. Springer Science & Business Media, 2012.

[5] INTERNATIONAL ORGANIZATION FOR STANDARDIZATION. *ISO 13374: Condition monitoring and diagnostics of machines — Data processing, communication and presentation*. Norma técnica, Genebra, 2015.

[6] INTERNATIONAL ORGANIZATION FOR STANDARDIZATION. *ISO 13379: Condition monitoring and diagnostics of machines — Data interpretation and diagnostics techniques*. Norma técnica, Genebra, 2015.

[7] INTERNATIONAL ORGANIZATION FOR STANDARDIZATION. *ISO 13381: Condition monitoring and diagnostics of machines — Prognostics*. Norma técnica, Genebra, 2015.

[8] INTERNATIONAL ORGANIZATION FOR STANDARDIZATION. *ISO 14224: Petroleum, petrochemical and natural gas industries — Collection and exchange of reliability and maintenance data for equipment*. Norma técnica, Genebra, 2006.

[9] INTERNATIONAL ORGANIZATION FOR STANDARDIZATION. *ISO 17359: Condition monitoring and diagnostics of machines*. Norma técnica, Genebra, 2018.

[10] INTERNATIONAL ORGANIZATION FOR STANDARDIZATION. *ISO 19860: Gas turbines — Data acquisition and trend monitoring system requirements for gas turbine installations*. Norma técnica, Genebra, 2005.

[11] SEIFFERT, C., KHOSHGOFTAAR, T. M., VAN HULSE, J., et al. "RUSBoost: A hybrid approach to alleviating class imbalance", *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, v. 40, n. 1, pp. 185-197, Out. 2010.

[12] KRSTAJIC, D., BUTUROVIC, L. J., LEAHY, D. E., et al. "Cross-validation pitfalls when selecting and assessing regression and classification models", *Journal of Cheminformatics*, v. 6, n. 1, pp. 1-10, Disponível em: <https://doi.org/10.1186/1758-2946-6-102014>. Acesso em: Julho de 2017.

[13] MALHI, A., GAO, R. X. "PCA-Based Feature Selection Scheme for Machine Defect Classification", *IEEE TRANSACTIONS ON INSTRUMENTATION AND MEASUREMENT*, v. 53, n. 6, pp. 1517-1525, Dez. 2004.

[14] LU, Y., COHEN, I., ZHOU, X. S., et al. "Feature Selection Using Principal Feature Analysis". In: *Proceedings of the 15th International Conference on Multimedia 2007*, pp. 301-304, Augsburg, Out. 2007.

[15] WANG, L., LEI, Y., ZENG, Y., et al. "Principal Feature Analysis: A Multivariate Feature Selection Method for fMRI Data", *Computational and Mathematical Methods in Medicine*, v. 2013, pp. 1-7, Disponível em: <http://dx.doi.org/10.1155/2013/645921>. Acesso em: Julho de 2017.

[16] JOLLIFFE, I. T. *Principal component analysis*. 2 ed. New York, Springer-Verlag, 2002.

[17] THEODORIDIS, S., KOUTROUMBAS, K. "Feature Generation I: Data Transformation and Dimensionality Reduction". In: *Pattern Recognition*, 4 ed., cap. 6, Academic Press, 2009.

[18] CADIMA, J. F. C. L., JOLLIFFE, I. T. "Variable selection and the interpretation of principal subspaces", *Journal of Agricultural, Biological, and Environmental Statistics*, v. 6, n. 1, pp. 1-62, Disponível em: <https://doi.org/10.1198/108571101300325256>. Acesso em: Julho de 2017.

[19] POWERS, D. M. W. "Evaluation: from Precision, Recall and F-measure to ROC, Informedness, Markedness and Corrrelation", *Journal of Machine Learning Technologies*, v. 2, n. 1, pp. 37-63, Disponível em: <http://hdl.handle.net/2328/27165>. Acesso em: Julho de 2017.

[20] SNOW, R., JURAFSKY, D., NG, A. Y. "Learning Syntactic Patterns for Automatic Hypernym Discovery". In: Saul, L. K., Weiss, Y., Bottou, L. (Eds.), *Advances in Neural Information Processing Systems 17*, MIT Press, pp. 1297-1304, 2005.

[21] SCHÖLKOPF, B., SMOLA, A., MÜLLER, K. R. "Nonlinear component analysis as a kernel eigenvalue problem", *Neural computation*, v. 10, n. 5, pp. 1299-1319, Jul. 1998.

[22] HOFFMANN, H. "Kernel PCA for novelty detection", *Pattern Recognition*, v. 40, n. 3, pp. 863-874, Set. 2007.

[23] AIZERMAN, M. "Theoretical foundations of the potential function method in pattern recognition learning", *Automation and remote control*, v. 25, pp. 821-837, 1964.

[24] BOSER, B. E., GUYON, I. M., VAPNIK, V. N. "A training algorithm for optimal margin classifiers". In: *Proceedings of the fifth annual workshop on Computational learning theory*, pp. 144-152. ACM, Jul. 1992.

[25] GENTON, M. G. "Classes of kernels for machine learning: a statistics perspective", *Journal of machine learning research*, v. 2, pp. 299-312, Dez. 2001.

[26] SHAWE-TAYLOR, J., CRISTIANINI, N. *Kernel methods for pattern analysis*. Cambridge university press, 2004.

[27] VAPNIK, V. *The nature of statistical learning theory*. Springer science & business media, 2013.

[28] HOFMANN, T., SCHÖLKOPF, B., SMOLA, A. J. "Kernel methods in machine learning", *The annals of statistics*, pp. 1171-1220, Jun. 2008.

[29] CHEN, B., LIANG, J., ZHENG, N., et al. "Kernel least mean square with adaptive kernel size", *Neurocomputing*, v. 191, pp. 95-106, Maio 2016.

[30] KOHLERT, M., KÖNIG, A. "Large, high-dimensional, heterogeneous multisensor data analysis approach for process yield optimization in polymer film industry", *Neural Computing and Applications*, v. 26, n. 3, pp. 581-588, Jul. 2014.

[31] MARTÍ, L., SANCHEZ-PI, N., MOLINA, J. M., et al. "Anomaly detection based on sensor data in petroleum industry applications", *Sensors*, v. 15, n. 2, pp. 2774-2797, Jan. 2015.