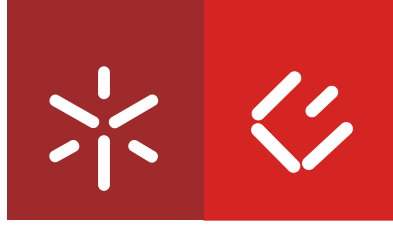




Universidade do Minho
Escola de Economia e Gestão

Stefan Santos Maciel Silva

**Desenvolvimento de um Modelo para Apoio
à Tomada de Decisão na Gestão dos Gastos
da Marinha do Brasil**



Universidade do Minho
Escola de Economia e Gestão

Stefan Santos Maciel Silva

**Desenvolvimento de um Modelo para Apoio
à Tomada de Decisão na Gestão dos Gastos
da Marinha do Brasil**

Proposta de Projeto de Mestrado
Mestrado em Estudos de Gestão

Trabalho efetuado sob a orientação do
Professor Doutor José Antônio Almeida Crispim

DIREITOS DE AUTOR E CONDIÇÕES DE UTILIZAÇÃO DO TRABALHO POR TERCEIROS

Este é um trabalho académico que pode ser utilizado por terceiros desde que respeitadas as regras e boas práticas internacionalmente aceites, no que concerne aos direitos de autor e direitos conexos.

Assim, o presente trabalho pode ser utilizado nos termos previstos na licença abaixo indicada.

Caso o utilizador necessite de permissão para poder fazer um uso do trabalho em condições não previstas no licenciamento indicado, deverá contactar o autor, através do RepositóriUM da Universidade do Minho.

Licença concedida aos utilizadores deste trabalho



Atribuição-NãoComercial-SemDerivações
CC BY-NC-ND

<https://creativecommons.org/licenses/by-nc-nd/4.0/>

DECLARAÇÃO DE INTEGRIDADE

Declaro ter atuado com integridade na elaboração do presente trabalho acadêmico e confirmo que não recorri à prática de plágio nem a qualquer forma de utilização indevida ou falsificação de informações ou resultados em nenhuma das etapas conducente à sua elaboração.

Mais declaro que conheço e que respeitei o Código de Conduta Ética da Universidade do Minho.

Desenvolvimento de um Modelo para Apoio à Tomada de Decisão na Gestão dos Gastos da Marinha do Brasil.

RESUMO

Em todo o mundo, governos enfrentam um desafio sem fim de promover um número cada vez maior de serviços públicos com o mesmo (ou até menor) volume de gastos a cada ano. Aprimorar a eficiência dos gastos no setor público significa potencializar a capacidade de atuação do estado, bem como garantir que o mesmo responda adequadamente àqueles que o financiam: seus cidadãos. Neste contexto, as tecnologias da informação e comunicação são a base para muitas das ideias inovadoras que permitem reduzir custos e, ao mesmo tempo, promover a qualidade do serviço prestado. No entanto, a literatura de inovação no setor público aponta que pouca atenção foi dada às inovações relacionadas a processos tecnológicos.

Por ser um campo de estudo com potencial de facilitar a tomada de decisão orientada a dados, esta pesquisa se propôs a efetuar uma abordagem baseada em aprendizado de máquina para o desenvolvimento de um modelo capaz apoiar o processo de tomada de decisão ao perceber a variação dos gastos comuns e essenciais nas mais de 400 instalações pertencentes à Marinha do Brasil, responsáveis pela execução de diversas e distintas atividades necessárias ao cumprimento de sua missão institucional em todo o território brasileiro. Para atingir este objetivo, dados secundários, oriundos de diferentes fontes e que representam diversos atributos das instalações, foram devidamente coletados, tratados, e carregados em uma arquitetura que possibilitou o processo de aprendizagem sob diversas perspectivas. Posteriormente, os melhores modelos passaram por um procedimento capaz de identificar os atributos mais relevantes com base na força com que os mesmos contribuem para a performance preditiva do modelo.

Os resultados mostram que, apesar das limitações de dados, o modelo alcançou uma capacidade razoável de previsão e, o conhecimento dos atributos mais importantes que contribuem para o mesmo são capazes de gerar percepções úteis que auxiliam no processo de tomada de decisão relacionado à gestão dos seus recursos. O desempenho alcançado também sustenta que a arquitetura desenvolvida também pode ser considerada um resultado adicional deste estudo. A implementação da mesma como um processo na plataforma de análise de dados da organização possibilitará uma análise mais aprofundada do problema apresentado, com um volume maior de dados e atributos, bem como sua reutilização em outras áreas de interesse.

Palavras-Chave: aprendizado de máquina; gastos militares; gestão de recursos; organização orientada a dados; tomada de decisão.

Development of a Model to Support Decision Making in the Management of Brazilian Navy Spending.

ABSTRACT

All around the world, governments face an endless challenge to promote an increasing number of public services with the same (or even less) volume of spending each fiscal year. Improving the efficiency of public sector spending means enhancing the state's capacity to act, as well as ensuring that it responds adequately to those who finance it: its citizens. In this context, information and communication technologies are the foundations for many of the innovative ideas that allow to reduce costs and, at the same time, promote the quality of the service provided. However, the innovation literature in the public sector points out that little attention was paid to innovations related to technological processes.

As it is a field of study with the potential to facilitate data-driven decision-making, this research proposed to carry out an approach based on machine learning for the development of a model capable of supporting the decision-making process by understanding the variation of common and essential expenses in the more than 400 Brazilian Navy installations, responsible for the execution of several and different activities necessary to fulfill its institutional mission throughout the Brazilian territory. In order to achieve this objective, secondary data, coming from different sources and representing different attributes of the facilities, were adequately collected, treated, and loaded into an architecture that enabled the learning process from different perspectives. Subsequently, the best models went through a procedure capable of identifying the most relevant attributes based on the strength with which they contribute to the predictive performance of the model.

The results show that, despite the data limitations, the model has achieved a reasonable capacity for forecasting, and the knowledge of the most important attributes that contribute to it are capable of generating useful insights that assist in the managerial decision-making process related to the allocation of resources. The performance achieved also maintains that the developed architecture can also be considered an additional result of this study. Its implementation as a process on the organization's data analysis platform will enable a more in-depth analysis of the problem presented, with a larger volume of data and attributes, as well as its reuse in other areas of interest.

Keywords: data-driven organization; decision making; machine learning; military spending; resource management.

ÍNDICE GERAL

RESUMO	iv
ABSTRACT	v
ÍNDICE GERAL	vi
ÍNDICE DE TABELAS	ix
ÍNDICE DE FIGURAS	x
LISTA DE SIGLAS E ABREVIATURAS	xi
CAPÍTULO 1 INTRODUÇÃO	1
1.1 Enquadramento do Problema e Relevância do Tema	2
1.2 Objetivos do Estudo	6
1.3 Contributos Esperados	7
1.4 Estrutura da Pesquisa	7
CAPÍTULO 2 REVISÃO DA LITERATURA	8
2.1 Introdução.....	8
2.2 Conceitos	8
2.2.1 E-Government	8
2.2.2 Data-Driven Decision Making.....	9
2.2.3 Capacidade Analítica das Organizações	10
2.2.4 Dados nos Estudos Organizacionais.....	12
2.2.5 Inteligência Artificial, Aprendizado de Máquina e Aprendizado Profundo.....	12
2.3 Relevância do Tema e Processo de Coleta	15
2.4 Conclusão	16
CAPÍTULO 3 METODOLOGIA.....	18
3.1 Introdução.....	18
3.2 Características Metodológicas.....	18
3.3 Aprendizado de Máquina: Potencialidades e Limitações	19
3.4 Abordagem Conceitual	21
3.4.1 Fase 1 – Extração	23
3.4.2 Fase 1 – Transformação.....	25
3.4.3 Fase 1 – Carga	26
3.4.4 Fase 2 – Entrada de Dados	27

3.4.5 Fase 2 – Atributos Categóricos	29
3.4.6 Fase 2 – Valores Ausentes.....	29
3.4.7 Fase 2 – Seleção de Atributo Alvo (AA)	31
3.4.8 Fase 2 – Particionamento de Dados	31
3.4.9 Fase 2 – Seleção de Atributos	33
3.4.10 Fase 2 – Identificação de <i>Outliers</i>	34
3.4.11 Fase 2 – Treinamento de Modelos: Pré-Processamento	35
3.4.12 Fase 2 – Treinamento de Modelos: Escolha e Configuração	36
3.4.13 Fase 2 – H2O <i>Automatic Machine Learning</i>	39
3.4.14 Fase 2 – Avaliação dos Resultados	40
3.5 Conclusão	42
CAPÍTULO 4 APRESENTAÇÃO E DISCUSSÃO DOS RESULTADOS	42
4.1 Introdução.....	42
4.2 Resultados do Treinamento Manual de Modelos	43
4.3 Resultados da Plataforma de Aprendizado Automático (H2O AutoML)	45
4.4 Discussão dos Resultados Consolidados.....	46
4.5 Resultados adicionais por meio de <i>Interpretable Machine Learning (IML)</i>	48
4.5.1 Importância dos Atributos por Permutação (<i>Permutation Feature Importance</i>).....	49
4.5.2 <i>Partial Dependence Plot (PDP)</i>	51
4.6 Conclusão	55
CAPÍTULO 5 SUGESTÃO DE IMPLEMENTAÇÃO	57
5.1 Introdução.....	57
5.2 Protótipo da <i>Framework</i> de Aprendizado.....	57
5.3 Fontes de Dados e Extração.....	58
5.4 Transformação e Carga.....	59
5.5 Banco de Dados Analítico e Apresentação	59
5.6 Potenciais Benefícios	60
CAPÍTULO 6 CONCLUSÕES.....	62
6.1 Contribuições da Pesquisa.....	64
6.2 Limitações do Estudo.....	65
6.3 Sugestões para Futuras Pesquisas.....	65
REFERÊNCIAS BIBLIOGRÁFICAS.....	67

APÊNDICE I – Relação de despesas enquadradas em Funcionamento (FC)	75
APÊNDICE II – Relação de despesas enquadradas em Manutenção (MN)	76
APÊNDICE III – Estrutura Organizacional da Marinha do Brasil.....	77

ÍNDICE DE TABELAS

Tabela 1 - Descrição das Fontes de Dados Externas	23
Tabela 2 - Identificação dos atributos obtidos.....	24
Tabela 3 - Procedimentos efetuados na etapa de transformação dos dados	25
Tabela 4 - Descrição dos Atributos-Alvo (AA) definidos	29
Tabela 5 - Performance dos Modelos Manuais para Despesas FC em P1 e P2.....	44
Tabela 6 - Performance dos Modelos Manuais para Despesas MN em P1 e P2	44
Tabela 7 - Performance dos Modelos H2O AutoML para as Despesas FC em P1 e P2	45
Tabela 8 - Performance dos Modelos H2O AutoML para as Despesas MN em P1 e P2	45
Tabela 9 - Comparativo de Performance dos Modelos para Despesas FC em P1 e P2	46
Tabela 10 - Comparativo de Performance dos Modelos para Despesas MN em P1 e P2	46
Tabela 11 - Informações sobre os modelos de melhor performance	56

ÍNDICE DE FIGURAS

Figura 1 - Vantagem Competitiva e Sofisticação Analítica. Adaptado de Competing on analytics: Updated, with a new introduction: The new science of winning, por Davenport, T., e Harris, J., 2017, Boston: Harvard Business Press. Direitos autorais 2017 por Harvard Business School Publishing Corporation.....	11
Figura 2 - Ilustração de um modelo de Aprendizado Profundo. Adaptado de Deep Learning (p. 6), por Goodfellow, I., Bengio, Y., & Courville, A., 2016, MIT press	13
Figura 3 - Resultados da pesquisa na Web of Science, com as palavras-chave “machine learning” e “decision making”, por área de pesquisa. Pesquisa atualizada em 18/02/2020.....	14
Figura 4 - Resultados da pesquisa na Web of Science, com as palavras-chave “machine learning” e “decision making”, por ano de publicação. Pesquisa atualizada em 18/02/2020.	14
Figura 5 - Atribuição e funcionalidade de um modelo preditivo.....	20
Figura 6 - Ilustração da Fase 1 da Abordagem Conceitual.....	22
Figura 7 - Ilustração da Fase 2 da Abordagem Conceitual.....	22
Figura 8 - Diagrama Entidade-Relacionamento.....	27
Figura 9 - Relacionamento entre as categorias de FNCD com a criação de perspectivas.	28
Figura 10 - Exemplo de one-hot encoding em variável categórica representando cores	29
Figura 11 - Análise dos valores ausentes em P1-FC realizado pelo pacote “mice”	30
Figura 12 - Análise dos valores imputados em P1-FC.....	31
Figura 13 - Exemplo gráfico de underfit e overfit no ajuste de modelos	32
Figura 14 - Representação gráfica do método k-folds cross-validation	33
Figura 15 - Análise da seleção de atributos realizado pelo pacote "Boruta".....	34
Figura 16 - Histograma do atributo ISPP antes e depois da transformação.	36
Figura 17 - Representação de uma Árvore de Decisão e suas consequências nos dados	38
Figura 18 - Funcionalidade "leaderboard" do H2O AutoML.....	40
Figura 19 - Observações Reais x Valores Previstos para P1-FC.....	48
Figura 20 - Importância dos Atributos nos Modelos da Perspectiva P1.....	49
Figura 21 - Resultados da técnica PDP nos atributos relevantes de P1-FC	52
Figura 22 - Resultados da técnica PDP nos atributos relevantes de P1-MN	53
Figura 23 - PDP de ISPP e FNCD sobre o as previsões do AA (FNCT)	56
Figura 24 - Ilustração de Processo ETL.....	58
Figura 25 - Implementação da rotina para Aprendizado de Máquina.....	60

LISTA DE SIGLAS E ABREVIATURAS

AA	Atributo-Alvo
AI	Artificial Intelligence
ANN	Artificial Neural Networks
AutoML	Automatic Machine Learning
DDDM	Data-Driven Decision Making
DER	Diagrama Entidade-Relacionamento
DL	Deep Learning
ETL	Extract Transform Load
FC	Funcionamento
GBM	Gradient Boosting Machines
IML	Interpretable Machine Learning
MAE	Mean Absolute Error
MB	Marinha do Brasil
ML	Machine Learning
MN	Manutenção
ND	Natureza de Despesa
OE	Objetivo Específico
OM	Organização Militar
PD	Plano Diretor
PDP	Partial Dependence Plot
RF	Random Forest
RLM	Regressão Linear Múltipla
RMSE	Root Mean Square Error
SIAFI	Sistema Integrado de Administração Financeira
SVR	Support Vector Regression
TIC	Tecnologia da Informação e Comunicação
XGB	XGBoost

CAPÍTULO 1 INTRODUÇÃO

Uma eficiente gestão dos recursos públicos, em qualquer área de atuação, sempre foi considerada fundamental por ter relação direta com a correta implementação, por um determinado Estado, de políticas públicas essenciais ao seu povo. Uma das razões para esta relevância é exposta por Ferreira e Oliveira (2017), ao assinalar que a adequada gestão orçamentária é a atividade pela qual faz-se o uso responsável dos recursos públicos para concretizar serviços e investimentos sociais que, em última instância, são encarados como uma simples contrapartida ao financiamento da atividade estatal pelo seus cidadãos.

Considerando a existência de um limite na capacidade dos governos em extrair recursos da sociedade para fazer frente a uma demanda cada vez maior por serviços públicos e investimentos, torna-se fundamental a racionalização e priorização do gasto público. Sua aplicação, além de ser transparente, deve também produzir resultados positivos que compensem o custo da retirada de recursos de seus cidadãos através da tributação (Boueri, Rocha, & Rodopoulos, 2015).

Uma vez inseridos em um cenário com recursos limitados e orçamentos reduzidos, a inovação na administração pública acaba por ser uma das respostas para reduzir custos e continuar a estimular o crescimento da economia. Neste contexto, Janssen e Estevez (2013) apontam algumas implementações que contribuem para alcançar este objetivo: abertura de dados governamentais, padronização de processos, fornecimento de serviços exclusivamente online, participação em mídias sociais, participação dos cidadãos em processos legislativos e decisões, entre outras.

Em um sistemático estudo da literatura existente sobre inovação no setor público, de Vries, Bekkers, e Tummers (2016) categorizaram 47% dos estudos como do tipo “inovações em processos”, dos quais apenas 15% se referem a inovações de caráter tecnológico, como por exemplo o movimento *e-government*, que trata do uso estratégico da Tecnologia da Informação e Comunicação (TIC) no âmbito do setor público (Bekkers & Homburg, 2005).

Entretanto, é certo que a tecnologia assume um papel fundamental nos processos de gestão atuais. Atualmente vivemos em uma era de excesso de dados, o que torna a identificação e separação do que é útil e relevante uma importante habilidade para a gestão. Na maioria dos casos, não faltam dados para a tomada de decisão, o que falta é a capacidade de gerenciá-los e explorá-los devidamente, o que pode ser entendido como a fundação, no âmbito das organizações (governos, empresas, entidades, etc.) para iniciar um processo de transformação que tem como objetivo orientar, de forma efetiva, a decisão com base em dados (Kiron, 2017). Apesar de ser um campo de estudo pouco

explorado no setor público, os benefícios advindos de inovações tecnológicas como *big data*, inteligência artificial e aprendizado de máquina já iniciaram um período de transição para um setor público orientado a dados, que almeja obter uma maior transparência e eficiência em termos de recursos utilizados (Agbozo & Asamoah, 2019).

Portanto, a razão principal que justifica o presente estudo reside no fato de que as modernas tecnologias de gerenciamento e análise de dados possuem o potencial de contribuir, se bem aplicadas, com o aprimoramento das atividades de planejamento e controle do orçamento, bem como no processo de tomada de decisão sobre sua aplicação.

1.1 Enquadramento do Problema e Relevância do Tema

Para que este trabalho seja efetivamente orientado para uma problemática concreta e bem definida, será necessária uma breve contextualização sobre o ambiente em que a Marinha do Brasil (MB) se insere, onde seus gastos são registrados e quais deles serão objeto do presente estudo, bem como os desafios que o modelo a ser proposto pretende superar. Desta forma, evita-se quaisquer desvios dos objetivos principais deste estudo sem deixar de fornecer informações suficientes para um sólido entendimento.

Primeiramente, é relevante saber que o Brasil possui a maior bacia hidrográfica do planeta, com 4,5 milhões de km² de área marítima e um litoral de 7400km de extensão que, recentemente, ganhou uma nova dimensão de importância com a descoberta e exploração de jazidas de petróleo em águas profundas. Tal dimensão territorial faz com que a Marinha Brasileira conte com mais de 70.000 militares e servidores civis, distribuídos em mais de 400 Organizações Militares (OM) espalhadas por todo o território brasileiro (Marinha do Brasil [MB], 2019; Ministério da Defesa, 2019b).

Dentre suas atribuições, além da proteção, monitoramento e controle do litoral brasileiro (que já é conhecida e não necessita detalhamento), outras ações subsidiárias são realizadas pela MB de forma permanente, como por exemplo, na lista exemplificativa constante no site do Ministério da Defesa:

- (...)
- Fiscaliza o cumprimento da legislação da segurança do tráfego aquaviário;
- Controla o ensino profissional marítimo e portuário, contribuindo para a formação de profissionais da Marinha Mercante em todo o País;
- Por meio do Sistema de Segurança do Tráfego Aquaviário (capitanias dos portos, delegacias e agências), zela pela salva-guarda da vida humana no mar e nas águas interiores, pela segurança da navegação e pelo controle da poluição hídrica por embarcações, plataformas e instalações de apoio;

- Por meio dos Navios de Assistência Hospitalar (NAsH), atende as populações carentes da Amazônia e do Pantanal Mato-Grossense. São as chamadas Operações de Assistência Hospitalar à População Ribeirinha (ASSHOP). Os NAsH são conhecidos pelas populações ribeirinhas como os “navios da esperança”;
- Atua na execução de Ações Cívico-Sociais (ACISO) em diversas comunidades carentes do País, por meio, entre outros, da recuperação de escolas e abrigos, da assistência médico-odontológica e da doação de sangue a entidades locais;
- Participa de operações de apoio humanitário, com a montagem de hospitais de campanha, inclusive no exterior (Ministério da Defesa, 2019a).

Com relação ao seu orçamento, a MB dispõe de um instrumento próprio para a gestão orçamentária e financeira, chamado Plano Diretor (PD), formalizado por meio de documentos internos, a seguir:

O Plano Diretor (PD) é um instrumento de planejamento, execução e controle, de caráter permanente, inerente às gestões orçamentária e financeira, desenvolvidas nos diversos escalões administrativos, visando à adequação dos recursos disponíveis às necessidades da Marinha do Brasil. (...) O Sistema do Plano Diretor (SPD) se constitui em um conjunto de conceitos, processos, regras de funcionamento, atores e procedimentos, os quais permitem o planejamento, a execução e o acompanhamento das atividades orçamentárias, bem como a produção de informações gerenciais necessárias à tomada de decisão (MB, 2014, p. 2-1)¹.

Além de ser essencial à gestão orçamentária da MB, ressalta-se os outros propósitos relevantes do PD, tais como permitir a aplicação de recursos com eficiência, eficácia, efetividade e economicidade, propiciar continuidade administrativa no que tange a gestão dos recursos disponíveis, bem como maximizar o atendimento das demandas internas (MB, 2014) .

Com o propósito de manter um aprimoramento constante de seus instrumentos, no ano de 2018 foi constituído um grupo de trabalho interno com o objetivo de aprimorar o SPD, por meio da avaliação da adoção de práticas e métricas focadas no orçamento por resultados, aprimoramento da gestão de despesas de custeio, bem como integração de seus processos ao Planejamento Estratégico da Marinha. Mais precisamente na gestão de despesas de custeio, deseja-se o estabelecimento de um critério único e padronizado de execução e avaliação das despesas comuns às OM e essenciais à sua operação diária. Tais despesas são classificadas, internamente, em duas categorias a saber: Funcionamento (FC) e Manutenção (MN).

¹ O documento que foi referenciado possui a seguinte estrutura de numeração de páginas: capítulo – número da página dentro do capítulo. Documento disponibilizado na intranet da organização e não está disponível publicamente.

A padronização que se deseja alcançar tem como fundamento algumas características relevantes da MB, que vão além da grande dimensão da organização, já evidenciada anteriormente. Para cumprir sua missão institucional, sua estrutura organizacional contempla instalações com as mais diversas atribuições (ensino, centros de instrução e adestramento, hospitais, policlínicas, depósitos, diretorias especializadas, bases operativas, instalações para reparo e manutenção de meios, desenvolvimento tecnológico, etc.), o que garante características distintas na forma e dimensão da aplicação destes recursos. A presente diversidade de instalações aliadas a vasta extensão territorial da organização resultou na aplicação, por gestores distintos, de metodologias diferentes para se aferir quais despesas são enquadráveis como FC e MN.

A situação exposta impede a formulação de indicadores padronizados, que permitam a aferição de dados atinentes a distintas OM sob um mesmo parâmetro comum, respeitadas as particularidades cabíveis. Tal fato prejudica a comparabilidade de informações, necessária à tomada de decisões gerenciais (Szuster, Cardoso, Szuster, Szuster, & Szuster, 2009). A ausência de comparabilidade compromete, ainda, a eficácia na implementação das medidas de economia, elevando a quantidade de recursos a serem aplicados para esta finalidade.

Neste contexto, o primeiro dos resultados do estudo do Grupo de Trabalho, voltado à obtenção da comparabilidade de informações entre as OM, foi estabelecer uma lista permanente de despesas que se enquadram em FC e MN. Tal lista foi construída tendo como base o conceito de Natureza de Despesa (ND), que é definida como uma codificação padronizada de 8 dígitos com a finalidade de detalhar e categorizar o objeto de gasto (Secretaria do Orçamento Federal [SOF], 2019). Os apêndices I e II listam as despesas enquadradas em cada categoria, acompanhado dos últimos 4 dígitos de sua ND.

Assim, toda vez que algum recurso orçamentário é aplicado na compra de algum material ou contratação de algum serviço, é obrigatória a definição de uma ND diretamente no Sistema Integrado de Administração Financeira do Governo Federal (SIAFI)². É realizado também o registro de quem fez a aplicação (no caso em estudo, qual OM) e de quando (data de emissão do documento). Cabe ressaltar que tais dados não possuem sigilo algum, sendo divulgadas de forma aberta ao cidadão, em

² É o Sistema Integrado de Administração Financeira do Governo Federal que consiste no principal instrumento utilizado para registro, acompanhamento e controle da execução orçamentária, financeira e patrimonial do Governo Federal.

plataformas do governo³, por estarem incluídas em seu direito fundamental de acesso à informação por meio eletrônico e sem justificar qualquer pedido (Brasil, 2011).

O segundo resultado do Grupo de Trabalho considerado relevante para este estudo é a agregação da gestão das despesas de MN e FC, antes espalhadas por diversos gestores e setores da MB, em uma única gerência. Assim, pretende-se submeter a todas as OM a um único processo de monitoramento de eficiência e qualidade do gasto, evitando-se que haja desequilíbrio entre exigências provenientes de setores distintos.

Pretende-se, com a combinação dos dois resultados acima expostos, após sua devida implementação e maturação nos processos da MB, alcançar os seguintes benefícios:

- Aprimoramento da eficiência na gestão dos recursos essenciais a operação das OM, conseqüentemente ampliando a aplicação em investimentos;
- Estabelecimento de indicadores de comparação entre OM com o posterior incentivo a intercâmbios de boas práticas entre setores;
- Estabelecimento de políticas de bonificação decorrentes de bons desempenhos e economias existentes; e
- Pleno monitoramento de gasto e gestão de medidas de economia, conforme a necessidade, para toda a MB.

Finalmente, os benefícios acima listados, previstos pela MB, acabam por ser os pilares para a problemática que se deseja expor. Implantar os resultados do Grupo de Trabalho nos processos da MB significa desenvolver um modelo para apoio à tomada de decisão que possa se beneficiar da comparabilidade entre OM trazida pela padronização que foi estabelecida, mas sem deixar de considerar as especificidades operativas de cada uma.

É interessante ressaltar o fato de ser incomum uma organização possuir instalações com atividades tão distintas umas das outras (por exemplo, hospitais, escolas e bases) e, conseqüentemente, demandas significativamente distintas por recursos humanos, materiais e orçamentários para manter-se em operação. A grande maioria dos estudos relacionados a área de Defesa pesquisa tópicos relacionados a sua atividade fim (tecnologias militares, melhorias no combate, logística, entre outros). Tal fato torna escasso o número de estudos acadêmicos que priorizam

³ Uma das plataformas disponíveis é o Portal da Transparência, que pode ser acessado a qualquer momento em <http://www.portaltransparencia.gov.br/despesas>

entender e aprimorar a gestão do orçamento na área de Defesa, levando em consideração tais peculiaridades.

1.2 Objetivos do Estudo

Considerando a existência de uma problemática concreta a ser resolvida e também uma lacuna, tanto a nível institucional, no que se refere a falta de processos consistentes capazes de apoiar a tomada de decisão na alocação recursos entre instalações com atividades distintas, quanto a nível de literatura, tendo a em vista a escassez de estudos que considerem as peculiaridades da organização no aprimoramento da gestão, o seguinte questionamento de partida foi proposto:

Como uma ferramenta computacional poderia apoiar o processo de tomada de decisão na gestão centralizada das despesas de FC e MN da organização, levando em consideração as especificidades de suas instalações e a diversidade de suas tarefas?

A partir da presente questão, foi definido como objetivo geral da pesquisa: **Desenvolvimento de um modelo para apoio à tomada de decisão na gestão dos gastos da Marinha do Brasil.**

De maneira mais detalhada, e com a preocupação de atingir os benefícios esperados pela organização, ficam estabelecidos os seguintes objetivos específicos (OE):

- OE1 - Identificar, coletar e analisar os dados atualmente disponibilizados que representem atributos das instalações da organização e que possam auxiliar no desenvolvimento do modelo.
- OE2 - Desenvolver uma infraestrutura capaz de gerenciar os dados coletados de forma eficaz, possibilitando o relacionamento entre os mesmos, bem como diversos tratamentos e abordagens.
- OE3 - Validar um modelo, baseado em aprendizado de máquina, que possa explicar e prever substancialmente as variações nos gastos de FC e MN das instalações da MB, identificando suas potencialidades e limitações para futuros estudos.
- OE4 - Avaliar se o modelo pode ser implantado de forma sistêmica dentro da organização, com o apoio de ferramentas tecnológicas de código aberto (software livre), de preferência.

1.3 Contributos Esperados

Considerando os efeitos práticos deste estudo, no curto prazo pretende-se subsidiar a MB com informações úteis que auxiliem na tomada de decisão acerca da aplicação de recursos orçamentários em FC e MN. Tal subsídio significa construir conhecimento no sentido de entender os fatores responsáveis por significativa parte das variações dos gastos nestas categorias. Existe a possibilidade de que as decisões tomadas com base nos achados deste estudo possibilitem a direta economia de recursos públicos, para que sejam aplicados em outros fins.

No médio prazo, a implementação do modelo de forma sistêmica, ao ampliar as variáveis analisadas e as perspectivas a serem abordadas e possibilitar sua repetição em outros contextos, permite que a organização estude variações em outros atributos que sejam do interesse da mesma, sem necessidade de conhecimento prévio nas áreas de estudo relacionadas e sem a participação de especialistas na área de tecnologia.

Também podem ser identificados contributos com efeitos na teoria. As limitações do modelo a ser proposto podem ser utilizadas para, no longo prazo, produzir conhecimento e subsídios em futuros estudos relacionados a implantação de modelos de apoio a decisão em ambientes organizacionais. O comportamento do gasto identificado neste trabalho pode vir a ser similar à de outras organizações, podendo se tornar algo característico de um ambiente com particularidades similares a do presente estudo.

1.4 Estrutura da Pesquisa

Esta introdução se propôs a apresentar o enquadramento do problema a ser estudado, bem como a relevância da pesquisa e objetivos a serem alcançados; o CAPÍTULO 2 destina-se a revisão da literatura, onde serão explorados conceitos relevantes para um pleno entendimento do estudo realizado, bem como uma breve análise sobre o que a literatura possui de relevante no que se refere a utilização de aprendizado de máquina para apoio a tomada de decisão e estudos semelhantes relacionados à gestão do orçamento na área de Defesa; o CAPÍTULO 3 apresenta a metodologia adotada; o CAPÍTULO 4 apresenta e discute os resultados e a interpretação dos modelos encontrados; o CAPÍTULO 5 propõe sugestões para implementação da metodologia de forma sistêmica e; por fim, o CAPÍTULO 6 apresenta a conclusão do estudo, apontando as limitações identificadas e as oportunidades para a condução de futuras pesquisas na área.

CAPÍTULO 2 REVISÃO DA LITERATURA

2.1 Introdução

A revisão da literatura deve, principalmente, alargar o contexto apresentado na introdução, buscando suportar o objetivo do trabalho com pesquisas previamente realizadas. No presente estudo, a revisão foi fundamental para constatar as lacunas presentes na literatura bem como identificar conceitos e os melhores métodos para abordar o problema. Desta forma, este capítulo organiza-se da seguinte forma: na seção 2.2 apresentam-se estudos que se relacionam os principais conceitos abordados neste trabalho, de forma que fique clara a relevância dos mesmos na abordagem a ser efetuada; a seção 2.3 descreve o processo de procura por pesquisas na área de Defesa envolvendo inovações tecnológicas para gestão do orçamento e apoio a tomada de decisão; por fim, a seção 2.4 reúne os aspectos mais relevantes da revisão.

2.2 Conceitos

2.2.1 E-Government

Conforme aponta Bekkers e Homburg (2005), apesar do conceito de *e-government* estar geralmente associado com o grau de utilização, de forma estratégica, da TIC no âmbito do setor público, também deve ser entendido como uma reformulação das relações e iterações governamentais com seus *stakeholders*. As constantes evoluções tecnológicas na área, juntamente com o incessante desafio no atendimento das demandas da sociedade, foram determinantes na criação de distintas fases, não necessariamente excludentes, mas complementares, representadas por mudanças substanciais na percepção e no uso da tecnologia.

Janssen e Estevez (2013) e Millard (2018) estudaram a evolução ocorrida na área, identificando de forma estruturada as diferentes perspectivas utilizadas ao longo dos anos, a seguir:

- *Eletronic-Government* (e-Government): aplicação da TIC em processos já existentes, aprimorando as atividades do setor público.
- *Transformational-Government* (t-Government): necessidade de mudanças estruturais demandaram aplicação da TIC em processos de negócio internos e externos.
- *Lean Government* (l-Government): criado como uma forma de combater a crise financeira mundial de 2008. Governos se viram obrigados a diminuir seus gastos e o tamanho de sua administração, utilizando TIC para manter e até mesmo aprimorar sua capacidade de

resolver problemas da sociedade, em colaboração com cidadãos e empresas. É entendido como uma forma de “fazer mais com menos”.

- *Open-Government* (o-Government): fase mais recente que prevê a utilização da TIC como possibilitador da integração total entre diversos níveis do setor público e atores relevantes fora do governo. Compartilhamento de infraestruturas, dados, recursos, processos e ferramentas entre administrações distintas. Como um de seus componentes principais, esta fase requer recursos abertos, incluindo dados.

Relativamente ao o-Government, apesar do conceito enfrentar uma série de desafios de carácter técnico, político e até mesmo cultural, não é difícil entender como a atividade governamental pode se tornar muito mais eficiente e efetiva quando suas partes compartilham os recursos, ativos e dados que necessitam. Gestores individuais, ou em grupos, serão capazes de aplicar ferramentas de análise nos dados compartilhados para aprender boas práticas, identificando o que é positivo ou negativo por meio da comparação, hierarquização e simulação entre contextos e/ou estratégias similares (Millard, 2018).

Deste tópico conclui-se que, apesar da inovação tecnológica no setor público ser uma área que necessita de mais estudos (de Vries et al., 2016), a literatura já constatou um relacionamento estreito entre as atividades governamentais e a aplicação de ferramentas tecnológicas para os mais diversos fins, trazendo resultados positivos para a sociedade de uma forma generalizada.

2.2.2 Data-Driven Decision Making

Data-Driven Decision Making (DDDM) se refere a prática de basear suas decisões na análise de dados, ao invés de puramente na intuição. Entretanto, sua adoção não exclui outras abordagens, e organizações ao redor do mundo podem aplicá-la com maior ou menor intensidade (Provost & Fawcett, 2013). DDDM é especialmente útil para as organizações que tem acesso a uma grande quantidade de dados que se relacionam, refletindo sua performance ao longo do tempo (Morrel-Samuels, Francis, & Shucard, 2009). Kiron (2017) interpreta a recente adoção de ferramentas tecnológicas de análise de dados em larga escala, por setores públicos e privados, como a “revolução na gestão”. Trata-se do sucesso na utilização da tecnologia para aprimorar a qualidade e variedade de seus produtos e serviços, se relacionar com seus clientes em uma forma mais profunda e modificar sua forma de operação.

Entretanto, ao se observar tantas falhas em iniciativas em orientar-se a dados dentro de uma organização, seu estudo aponta como uma das causas a ausência da base onde todo o processo vai se apoiar: dados de qualidade e bem gerenciados. Somente após assegurar a periodicidade, a acurácia, a

completude e a confiança dos dados, é que a organização deve seguir a próxima etapa, que é tratar o dado com um ativo organizacional essencial para a estratégia. Essa integração entre dados e tomada de decisão pode ser considerado outro desafio, especialmente entre gestores que estão acostumados a decidir com base na experiência, intuição ou instinto (Kiron, 2017). Em sua pesquisa, Kumar et al. (2013) aponta que nem todas as organizações podem aplicar DDDM, sendo por não possuir dados suficientes sobre os clientes (29% dos entrevistados) ou por não possuir a capacidade de converter uma grande quantidade de dados em conhecimento aplicável (39% dos entrevistados).

Apesar das dificuldades de implementação, orientar-se a dados pode ser altamente benéfico para a gestão. Foi estudada a existência, por meio de uma medida que classifica organizações de acordo com a intensidade com que as mesmas utilizam dados para tomar decisões, de uma correlação positiva entre o processo decisório baseado em dados e um aumento da produtividade de 5% a 6%, sugerindo que essa capacidade pode ser considerada um ativo intangível da organização, que obviamente é valorizada por investidores e, conseqüentemente, contribui para o aumento da rentabilidade (Brynjolfsson, Hitt, & Kim, 2011). Já com relação ao setor público, inovações orientadas a dados interligados a partir de uma variedade de fontes distintas, podem resultar em significantes transformações e causar benefícios para toda a sociedade como redução da poluição, maior eficiência energética e menores custos além de aprimorar a tomada de decisão, adoção de políticas públicas e o gerenciamento de crises (Janssen, Konopnicki, Snowdon, & Ojo, 2017). As oportunidades advindas de um setor público orientado a dados também são estudados por Christodoulou et al. (2018), resultando no aumento da eficiência na tomada de decisão e dos serviços prestados, bem como mais participação e transparência nos processos, de forma a fortalecer o relacionamento entre governo e cidadãos.

2.2.3 Capacidade Analítica das Organizações

Davenport e Harris (2017) definem capacidade analítica como o uso extensivo de dados, análises estatísticas e quantitativas, modelos preditivos e explicativos, além de uma gestão baseada em fatos para guiar decisões e ações. Além disso, os pesquisadores expõem a categorização a seguir:

- Analítica Descritiva: Provê acesso a dados históricos e atuais para sua exploração, utilizando dados internos e externos de diversas fontes.
- Analítica Preditiva: Envolve o uso de técnicas quantitativas e tecnologias para usar dados do passado para prever o futuro.
- Analítica Prescritiva: Também baseada em técnicas quantitativas e tecnologias para identificar comportamentos e ações que levam a resultados ótimos.

- Analítica Autônoma: Envolve inteligência artificial para criar e aprimorar modelos que aprendem com os dados, com uma menor participação de analistas.

A Figura 1 relaciona tais categorizações e demonstra um potencial aumento em termos de vantagem competitiva de acordo com o grau de sofisticação analítica adotada.

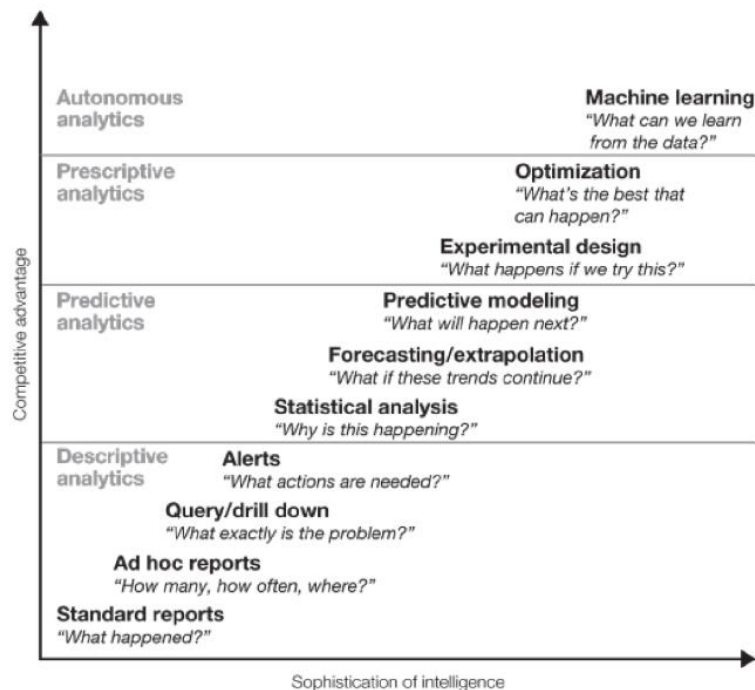


Figura 1 - Vantagem Competitiva e Sofisticação Analítica. Adaptado de Competing on analytics: Updated, with a new introduction: The new science of winning, por Davenport, T., e Harris, J., 2017, Boston: Harvard Business Press. Direitos autorais 2017 por Harvard Business School Publishing Corporation.

Anderson (2015) defende que diferentes partes da organização podem estar comprometidas em análises com também diferentes graus de sofisticação ao mesmo tempo e, portanto, a categorização exposta não deve ser entendida como uma progressão do nível de maturidade analítica das organizações, onde o próximo passo se inicia após ter finalizado o anterior. LaValle, Lesser, Shockley, Hopkins, e Kruschwitz (2011) efetuaram uma pesquisa que envolveu mais de 3.000 gerentes sobre o uso e crença no valor das capacidades analíticas e, ao separar as organizações com base na performance em sua área, concluiu que, comparativamente com as de pior performance, as organizações de melhor performance possuem:

- Probabilidade cinco vezes maior de utilizar alguma ferramenta analítica;
- Probabilidade duas vezes maior de utilizar alguma ferramenta analítica para guiar estratégias futuras;
- Probabilidade duas vezes maior de utilizar alguma ferramenta analítica para guiar as operações diárias;

- Probabilidade três vezes maior de possuir uma vasta experiência em sofisticadas ferramentas analíticas ao longo de uma ampla gama de funções, utilizando-as como um diferencial competitivo.

2.2.4 Dados nos Estudos Organizacionais

No estudo empírico de como indivíduos constroem estruturas e processos organizacionais, e como estes moldam e influenciam as relações entre pessoas, não existe consenso sobre o que realmente deve ser considerado um dado relevante. Atitudes, símbolos, documentos, sentimentos, significados e figuras podem ser coletados, estudados para apoiar ou refutar determinado posicionamento. Neste contexto, o estudo conduzido por Stablein (1999) define dados como representações que mantem uma correspondência bidirecional entre a realidade observada e um sistema simbólico. Em outras palavras, um dado de qualidade é aquele em que a codificação de uma observação em um sistema simbólico permite fazer inferências verdadeiras sobre a realidade. Desta forma, a correspondência bidirecional atua como um elemento fundamental para determinar a relevância de um determinado dado para estudos.

Em uma abordagem mais técnica, Anderson (2015), expõe os principais requisitos que os dados precisam possuir, além de serem relevantes, para que possam contribuir no processo organizacional de se orientar a dados:

- Devem representar o mesmo período temporal que se deseja pesquisar sobre;
- Devem ser precisos, limpos e imparciais;
- Deve ser originado de uma fonte confiável;
- Deve ser acessível de forma fácil, integrável com outros dados, e compartilhável;

2.2.5 Inteligência Artificial, Aprendizado de Máquina e Aprendizado Profundo

Aprendizado de Máquina (*Machine Learning* – ML) pode ser entendido como uma área de estudo, pertencente Inteligência Artificial (*Artificial Intelligence* – AI), que tem por objetivo utilizar diferentes técnicas computacionais para aprender e aprimorar-se, de forma autônoma, com base em dados. Enquanto ML trata especificamente no processo de aprendizado, AI foca-se, além do aprendizado, em resolução de problemas e raciocínio (El Naqa & Murphy, 2015; Kersting, 2018).

Goodfellow, Bengio, e Courville (2016) explicam que, nos primeiros estudos sobre AI, pesquisadores demonstraram que seria possível resolver, de forma rápida, problemas intelectualmente difíceis para os seres humanos, como complexos problemas de otimização combinatória. No entanto, o verdadeiro desafio residia em fornecer soluções para problemas que os humanos resolvem intuitivamente, como reconhecer o som de palavras ou rostos em imagens. Após várias tentativas, sem sucesso, que se propuseram a codificar o conhecimento necessário para resolver tais problemas, a solução encontrada foi permitir que os computadores obtivessem conhecimento, de forma autônoma, a partir da experiência, analisando os dados em termos de uma hierarquia de várias camadas de conceitos relacionados. A técnica recebeu o nome de Aprendizado Profundo (*Deep Learning* – DL) e é baseada em redes neurais (*Artificial Neural Networks* – ANN), que simula o funcionamento de cérebro, com neurônios e sinapses, para a solução de problemas. A Figura 2 ilustra, de forma conceitual, um modelo de Aprendizado Profundo.

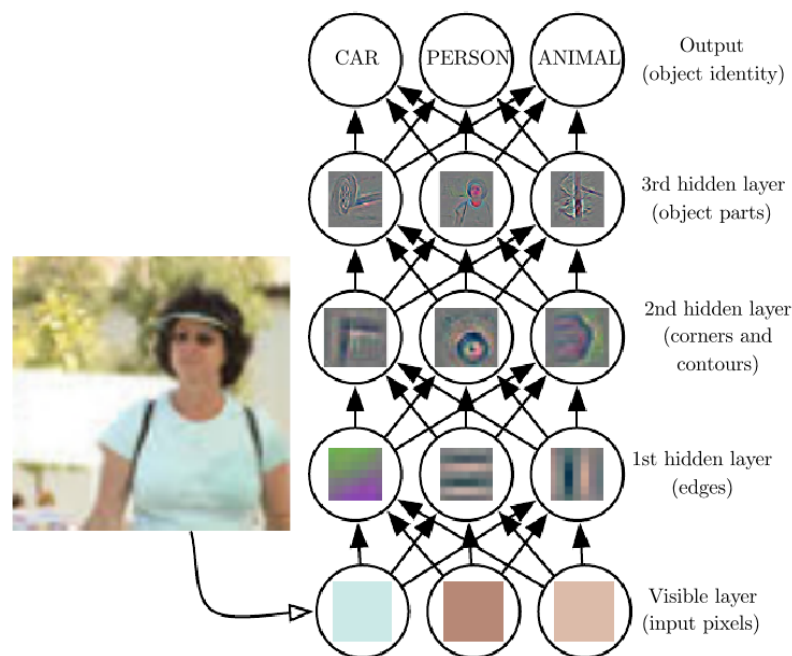


Figura 2 - Ilustração de um modelo de Aprendizado Profundo. Adaptado de Deep Learning (p. 6), por Goodfellow, I., Bengio, Y., & Courville, A., 2016, MIT press

A estrutura de um modelo de Aprendizado Profundo é baseada em camadas, onde cada uma representa a execução de um conjunto de instruções, e as camadas superiores podem se remeter aos resultados das anteriores. Quanto maior a profundidade das camadas, mais instruções são executadas e mais "detalhes" podem ser capturados.

De uma forma geral, as técnicas e algoritmos de ML necessitam de uma grande quantidade de dados para o processo de aprendizado e, portanto, se beneficiam do próprio contexto atual, onde, em

praticamente todas as áreas é gerado um volume imenso de dados nas mais diversas formas, estruturadas ou não. Abiodun et al (2018) e El Naqa & Murphy (2015) listam diversas áreas de estudo onde ML já foi aplicado com sucesso, como reconhecimento de padrões e de fala, sistemas de recomendação, visão computacional, engenharia, finanças, entretenimento, biologia e medicina.

Com o objetivo de identificar, de uma maneira abrangente, a relevância atual deste conceito e as principais áreas de estudo de sua aplicação, foi efetuada uma pesquisa na ferramenta de procura *Web of Science* utilizando as palavras-chave “machine learning” e “decision making”, de forma associada, sendo encontrados um total de 6.957 registros. A organização deste resultado por ano de publicação e por área de pesquisa são exibidos nas Figura 3 e Figura 4, respectivamente.

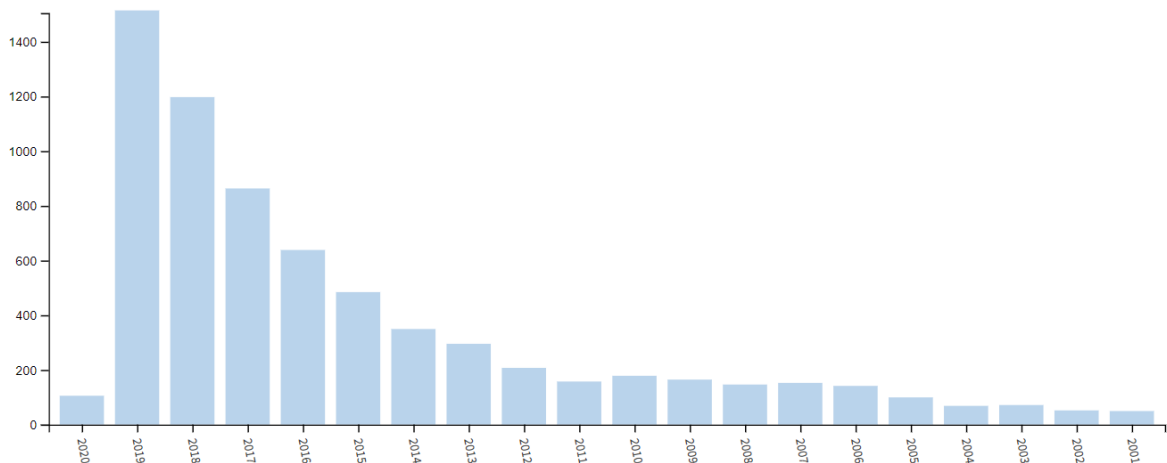


Figura 4 - Resultados da pesquisa na Web of Science, com as palavras-chave “machine learning” e “decision making”, por ano de publicação. Pesquisa atualizada em 18/02/2020.

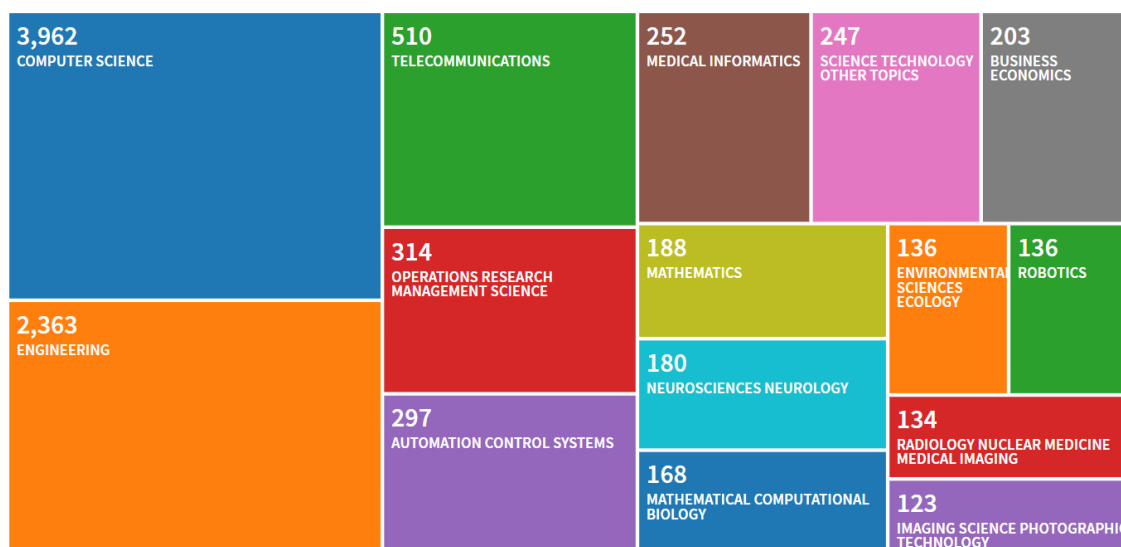


Figura 3 - Resultados da pesquisa na Web of Science, com as palavras-chave “machine learning” e “decision making”, por área de pesquisa. Pesquisa atualizada em 18/02/2020.

Dos resultados da pesquisa conclui-se que os últimos anos presenciaram um expressivo aumento nos estudos relacionados ao tema em questão, fruto da consolidação do aprendizado de máquina como uma ferramenta adicional para apoiar o processo de tomada de decisão. Com relação às áreas de estudo, apesar de ser possível identificar uma certa diversidade, é presente um foco na área de engenharia. De forma pontual, alguns dos estudos envolvem: modelo híbrido de ML e técnicas de otimização para alocação orçamentária em projetos de pesquisa e desenvolvimento (Jang, 2019), modelo híbrido de ML e técnicas de análise de decisão para classificação de inventários (Kartal, Oztekin, Gunasekaran, & Cebi, 2016), estimativa de margens de lucro em projetos de construção (Bilal & Oyedele, 2020), apoio a decisão na definição de tratamentos radioterápicos personalizados a cada paciente (Valdes et al., 2017), estimativas de consumo de energia (Edwards, New, & Parker, 2012; Robinson et al., 2017), modelos para previsão e resposta à demanda (Ghalekhondabi, Ardjmand, Young, & Weckman, 2017; Pallonetto, De Rosa, Milano, & Finn, 2019), previsão de produção agrícola (Chlingaryan, Sukkarieh, & Whelan, 2018), previsão da resistência à compressão de materiais (Yaseen et al., 2018), modelos para previsão de câncer (Xiao, Wu, Lin, & Zhao, 2018) e veículos autônomos (Ye et al., 2018).

2.3 Relevância do Tema e Processo de Coleta

Esta seção tem como objetivo analisar, na literatura, artigos que tratam do orçamento na área de Defesa e, dentre estes, aqueles que o analisam por alguma categoria de despesa, como os apresentados nos apêndices I e II.

Primeiramente, uma pesquisa na ferramenta de procura *Web of Science* utilizando as palavras-chave “defense OR defence” e “military”, de forma associada, retorna 10.407 resultados onde as áreas de estudo dominantes são Engenharia (14,43%), Leis Governamentais (14,40%), Relações Internacionais (14,10%). Nota-se um acentuado foco em tecnologias militares (Rath, Pattanayak, & Pati, 2016), aprimoramentos de combate (Sudhakar, Madhusudhan Reddy, & Srinivasa Rao, 2016), relações civil-militares e seu impacto ciência política (Feaver, 1999).

Posteriormente, ao acrescentar a pesquisa anterior os termos “spending OR expenditure OR budget”, os resultados são reduzidos para 1.190 resultados onde as áreas de estudo dominantes são Economia de Negócios (41,76%), Leis Governamentais (22,01%) e Relações Internacionais (20,75%). As despesas militares incluem pessoal militar e civil; operação e manutenção; compras; pesquisa e desenvolvimento; e ajuda militar (nas despesas militares do país doador) (Stockholm International Peace Research Institute [SIPRI], 2018). Na literatura, é possível encontrar diversos artigos e bancos de

dados sobre o gasto de organizações militares envolvendo: a) economia do país (proporção do PIB) (SIPRI - The World Bank, 2018), b) os impactos no crescimento da economia (Alptekin & Levine, 2012) e também na produtividade do país (Caruso & Francesco, 2012) e c) suas características determinantes (Hou, 2018).

Segundo De Rezende e Blackwell (2019), no Brasil (em 2018), cerca de 75% do orçamento de defesa são alocados em gastos com pessoal e previdência social, 12% em operação e manutenção e quase 13% em investimentos (aquisição e desenvolvimento de armas e outros equipamentos). A categoria de despesas de operação e manutenção representam os gastos com bens e serviços que sustentam as Forças Armadas (por exemplo, energia, alimentos, água, têxtil). Dada a rigidez dos gastos com pessoal, a manutenção e operação assume uma posição crítica na gestão de cada instalação militar. As restrições do orçamento nesta categoria podem influenciar diretamente a eficiência das forças militares, afetando a prontidão de combate obtida pelo treinamento (Svendsen, Kalita, & Gebhart, 2005).

Finalmente, ao combinar a pesquisa anterior com outros termos, como “decision making” (53 resultados), “expense category” (1 resultado) ou “water OR energy” (28 resultados), e analisando os artigos encontrados, não foi possível localizar pesquisas anteriores que tratam problema similar ao exposto neste estudo: uma única organização, composta por centenas de instalações com atividades bastante distintas, buscando uma maneira de obter um entendimento único e aprofundado das despesas comuns e essenciais que estão sendo feitas. O comportamento das despesas identificadas neste trabalho pode ser semelhante ao de outras organizações e tornar-se algo característico de um ambiente com as mesmas peculiaridades. Além disso, as limitações do modelo proposto podem ser usadas como pontos de partida de estudos futuros relacionados à implementação de modelos de suporte à tomada de decisão em ambientes organizacionais.

2.4 Conclusão

A seguir listam-se os aspectos mais relevantes encontrados por meio da revisão da literatura:

- A existência de um relacionamento estreito entre as atividades governamentais e a aplicação de ferramentas tecnológicas é capaz de gerar resultados positivos para a sociedade.
- Apesar das dificuldades de implementação, orientar-se a dados pode ser altamente benéfico para a gestão da organização e deve ser considerado um ativo intangível da

mesma. No setor público, essa cultura pode resultar em benefícios e significantes transformações.

- Empresas privadas líderes de mercado possuem maiores chances de possuírem vasta experiência em sofisticadas ferramentas analíticas para guiar estratégias futuras e as operações diárias.
- Pesquisas envolvendo a utilização de aprendizado de máquina para apoio a tomada de decisão, em diversas áreas de estudo, seguem em uma crescente nos últimos anos.
- Poucos estudos se propõem a apoiar a decisão na gestão do orçamento na área de Defesa e não foi possível encontrar estudos que o fazem por categoria de despesa.

CAPÍTULO 3 METODOLOGIA

3.1 Introdução

Este capítulo está estruturado da seguinte forma: a seção 3.2 contém as considerações metodológicas da pesquisa; a seção 3.3 trata de algumas potencialidades e limitações do Aprendizado de Máquina, bem como o que se espera atingir com o método; a seção 3.4 apresenta a modelagem conceitual desenvolvida para estudar o problema com uma detalhada descrição de todos os seus passos; por fim, na seção 3.5 é realizada uma breve conclusão sobre o capítulo. Vale ressaltar que, sempre que for relevante, serão identificadas as relações das abordagens efetuadas com os Objetivos Específicos (OE) apresentados na seção 1.2.

3.2 Características Metodológicas

Com o propósito de esclarecer as características metodológicas do presente estudo, as mesmas serão apresentadas conforme as definições e categorias expostas por Saunders, Lewis, e Thornhill (2016) e Singh (2007):

- **Filosofia:** Refere-se ao sistema de crenças e assunções adotadas para o desenvolvimento de conhecimento. Este estudo se reveste de uma filosofia positivista, com foco epistemológico em descobrir fatos e regularidades observáveis e mensuráveis por meio das relações existentes nos dados. Outra dimensão que auxilia no entendimento das filosofias de pesquisa são os paradigmas propostos por Burrell e Morgan (1979), onde o presente estudo se enquadra no paradigma funcionalista.
- **Raciocínio Lógico:** Refere-se à utilização da teoria relativamente ao processo de pesquisa. Este estudo possui uma abordagem indutiva, caracterizada pela coleta de dados para explorar um fenômeno, identificar padrões e criar, ao final, um modelo conceitual. Generaliza-se a partir do específico (o caso exposto neste trabalho) para o geral.
- **Escolhas metodológicas:** Primeiramente utiliza-se uma abordagem quantitativa, onde serão examinadas relações entre variáveis mensuradas numericamente e analisadas por métodos que envolvem técnicas estatísticas. Quanto aos objetivos, a natureza é exploratória-descritiva. Exploratória pelo propósito de querer descobrir o que ocorre relativamente a uma situação apresentada, de forma a esclarecer o entendimento de um determinado problema. No presente estudo, a parte exploratória reside na incerteza perante ao comportamento dos gastos de MN e FC da organização, o que levou a coleta abrangente de

dados de diferentes fontes (atributos) na tentativa de ganhar conhecimento e direcionar a pesquisa para os melhores resultados possíveis. Posteriormente, o estudo é também descritivo, a medida que busca mensurar o relacionamento entre os gastos e os atributos coletados, bem como identificar quais dos atributos coletados são os mais relevantes para uma melhor performance.

- **Estratégia:** Refere-se ao plano do pesquisador sobre como responder suas questões de pesquisa. O estudo caracteriza-se por um estudo de caso, que pode ser definido como uma investigação detalhada sobre a dinâmica de um determinado fenômeno dentro do contexto onde o mesmo ocorre.
- **Horizonte de Tempo:** A pesquisa é transversal, ou seja, estuda o problema em um período específico de tempo, já que o mesmo tem por base analisar os gastos das instalações da organização efetuados no ano de 2018.

3.3 Aprendizado de Máquina: Potencialidades e Limitações

Provost e Fawcett (2013) descrevem o Aprendizado de Máquina (ML) como uma subárea da Inteligência Artificial que tem como preocupação utilizar métodos de reconhecimento de padrões, matemáticos e estatística aplicada para aprimorar o conhecimento ou performance de um agente, conforme a sua experiência. Em outras palavras, um modelo decorrente deste método tende a analisar os dados oferecidos para fazer previsões antes inexistentes. Como previsão entende-se que o modelo é capaz de estimar um valor para a variável dependente que está sendo estudada (geralmente chamada de atributo alvo) com base nas variáveis independentes oferecidas como entrada (geralmente chamadas de atributos). A Figura 5 ilustra um exemplo bidimensional onde um modelo foi capaz de associar corretamente uma função matemática que relaciona os atributos observados com o atributo alvo, com margens de erro relativamente pequenas que possibilitam efetuar previsões.

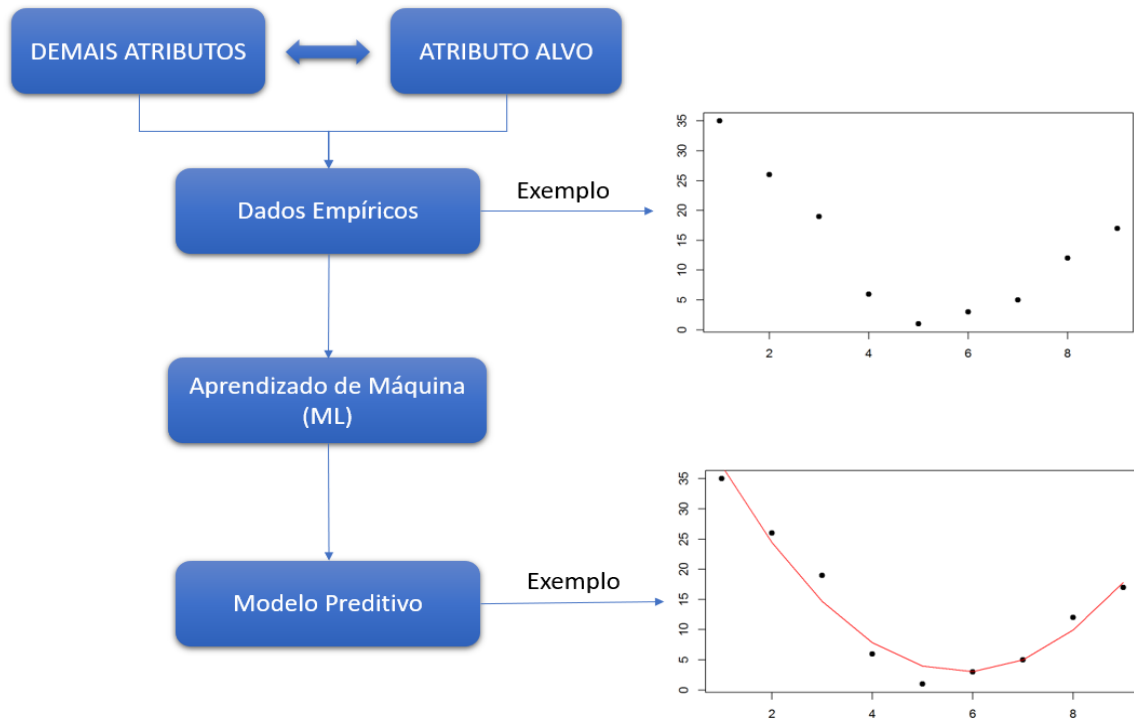


Figura 5 - Atribuição e funcionalidade de um modelo preditivo

Com base na figura exposta, este estudo considera como atributos os dados que representam as características das instalações da MB, e os seus respectivos gastos em MN/FC como atributo alvo. Espera-se, portanto, que o resultado seja um modelo capaz de efetuar previsões de gastos com base nas características dessas instalações, com o potencial de apoiar o processo de decisão dos gestores responsáveis pela alocação destes recursos em diversas situações distintas, como por exemplo para fornecer uma estimativa de gastos por ocasião da ativação de novas instalações ou em modificações de recursos em instalações já existentes, ou na identificação de quais atributos são os mais relevantes para a performance do modelo.

Entretanto, como qualquer outro método, ML possui diversas limitações, e a performance do modelo estará diretamente relacionada com uma série de fatores, como a qualidade dos dados, potencial de informação dos atributos, características do processo de aprendizado, entre outras. Lee, Yoo, Kim, Lee, e Hong (2019) explicam as dificuldades da utilização de técnicas baseadas em ML para a solução de problemas, e os principais aspectos estão resumidos a seguir:

- Identificação do problema: a tarefa que soluciona o problema deve ser facilmente relacionada a uma atividade de ML como regressões (previsões), classificação, agrupamento, entre outras.
- O processo de “limpeza”: é essencial selecionar ou extrair aqueles atributos que são realmente relevantes para a tarefa. No caso de os dados analisados serem originados de

fontes distintas, é necessário integrá-los de forma eficiente de forma a não perder informação.

- Seleção do algoritmo: aplicar o algoritmo correto pode ser considerado um dos maiores desafios, já que depende da qualidade dos dados, do seu tamanho, da tarefa envolvida e do tempo disponível para executá-la. Cada algoritmo possui vantagens e desvantagens que devem ser levadas em consideração.
- Configuração de parâmetros: alguns algoritmos necessitam ser configurados para alcançar melhores resultados. Tal configuração é alcançada ao selecionar valores adequados a seus parâmetros, o que requer profundo conhecimento do funcionamento do algoritmo e de suas características.

Além das dificuldades acima expostas, Du, Liu, e Hu (2019) apontam uma das mais significativas desvantagens de uma abordagem baseada em ML: a falta de transparência sobre o funcionamento de um modelo treinado, uma vez que seu comportamento não é simples de ser interpretado e, por várias vezes, seus usuários podem ficar sem saber como uma previsão específica foi calculada. Os autores também descrevem um vasto e recente campo de estudo chamado *Interpretable Machine Learning* (IML) que tem por objetivo dotar os modelos treinados com a habilidade de explicar ou apresentar seu próprio comportamento em termos fáceis de entender.

Uma série de técnicas, elencadas e descritas na seção 3.4 e também no Capítulo 4, serão aplicadas na tentativa de superar as dificuldades expostas acima e também dotar o modelo com algum potencial de interpretabilidade.

3.4 Abordagem Conceitual

Antes de apresentar os detalhes da abordagem deste estudo, é crucial descrever algumas premissas que foram consideradas durante o processo de desenvolvimento do mesmo. Em primeiro lugar, os recursos de software usado neste trabalho estão disponíveis gratuitamente na Internet, de forma que qualquer implementação futura baseada neste estudo (pela organização em foco ou outras) não implique em custos adicionais de licenciamento de software. Em segundo lugar, alguns recursos computacionais deste estudo foram idealizados para reutilização e flexibilidade, permitindo a operação em outros contextos, com dados referentes a outros problemas que a organização possa vir a ter ou mesmo na tentativa de melhorar os resultados encontrados nesta pesquisa. Finalmente, é importante ressaltar que a abordagem desenvolvida pode ser considerada relativamente simples, pois não tem

como foco obter um modelo final com a melhor performance possível, e sim verificar se ML pode, de fato, ser utilizado neste contexto para apoiar a tomada de decisão dos gestores.

A abordagem pode ser dividida em duas fases. A primeira fase guarda estrita relação com o OE1 e OE2 e envolve etapas relacionadas a coleta, tratamento e armazenamento dos dados obtidos. A segunda fase se relaciona com os OE3 e OE4 e é composto por um algoritmo, desenvolvido na linguagem R (R Core Team, 2017) que recebe os dados e aplica uma série de técnicas baseadas em ML para analisar os dados, treinar os modelos e avaliar os resultados. As Figuras 6 e 7 apresentam diagramas que ilustram as duas fases da abordagem conceitual, respectivamente, e as seções a seguir tem o propósito de detalhar cada etapa da mesma.

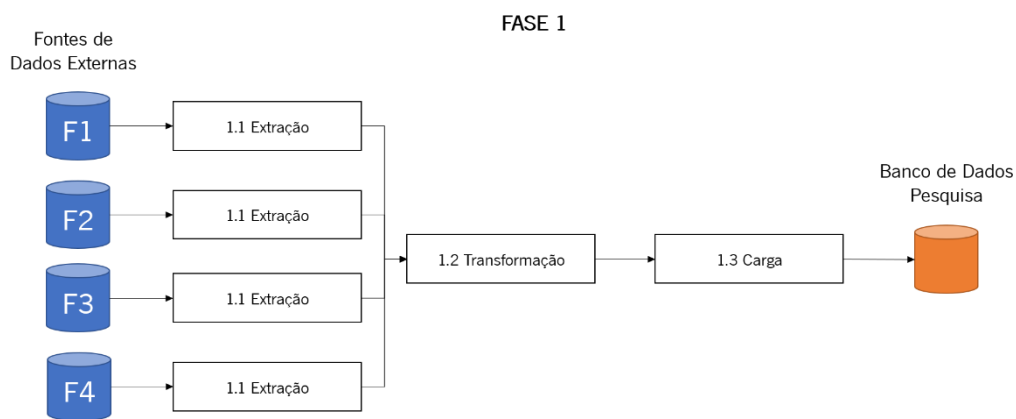


Figura 6 - Ilustração da Fase 1 da Abordagem Conceitual

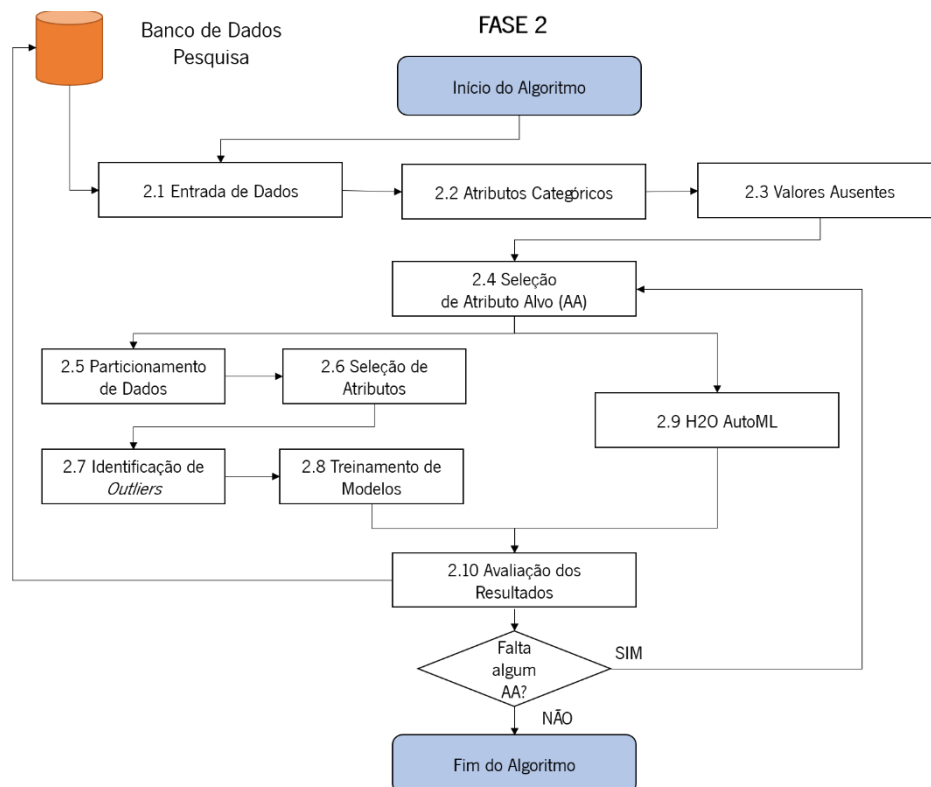


Figura 7 - Ilustração da Fase 2 da Abordagem Conceitual

3.4.1 Fase 1 – Extração

Esta etapa envolve o processo de identificação e coleta dos dados que representam diversas características das instalações da MB. Todos os dados utilizados neste estudo foram originalmente coletados para fins não relacionados à pesquisa e, portanto, quando se lida com dados secundários é necessária uma demonstração da sua efetiva utilidade nesse novo contexto (Stablein, 1999). De uma forma geral, todas as características possuem o potencial de explicar a variabilidade dos gastos MN/FC, seja por representarem a estrutura da instalação (quantidade de pessoas, atividade fim, nível hierárquico, etc.) seja por representarem características geográficas (condições climáticas) ou sociais (indicadores da cidade, estado ou região onde se localiza a instalação). Sobre a demonstração da utilidade dos dados secundários, o algoritmo presente na Fase 2 será responsável por esta verificação, por meio da avaliação da relevância dos atributos (o potencial que um atributo tem de explicar as variações sem implicar causalidade).

Esta etapa foi propositalmente repetida na Figura 6 para ilustrar que a extração ocorreu de forma distinta para cada fonte de dados externa utilizada. A Tabela 1 descreve as fontes externas utilizadas e uma breve descrição do processo de extração.

Tabela 1 - Descrição das Fontes de Dados Externas

Código	Nome da Fonte Externa	Descrição da Extração
F1	Plataforma de Análise Gerencial da MB	Consiste em uma plataforma analítica de informações orçamentárias e financeiras diversas sobre as instalações de toda a MB. A extração ocorreu de forma automática, por acesso direto ao Banco de Dados que armazenam os dados.
F2	Banco de Dados de Pessoal	Consiste em um banco de dados gerenciado pela diretoria responsável pelos recursos humanos da organização. A extração ocorreu mediante solicitação, sendo entregue uma planilha com os dados por instalação.
F3	Cadastro de Imóveis da Marinha	Consiste em um banco de dados gerenciado pela diretoria responsável pela gestão do patrimônio imobiliário da organização. A extração ocorreu mediante solicitação, sendo entregue uma planilha com os dados.
F4	Dados Cidades e Estados do Brasil (Instituto Brasileiro de Geografia e Estatística, 2019)	Consiste em informações disponibilizadas para consulta no sítio do Instituto Brasileiro de Geografia e Estatística (IBGE). Foi possível exportar os dados dos estados brasileiros em uma única planilha. Para os municípios, foi possível exportar uma planilha por estado brasileiro.
F5	Banco de Dados Meteorológicos para Ensino e Pesquisa (Instituto Nacional de Meteorologia -	Banco de dados que abriga séries históricas de estações meteorológicas espalhadas pelo país. Foi possível obter

	INMET, 2019)	uma planilha para cada estação meteorológica.
F6	Sistema Integrado de Administração Financeira (SIAFI)	Consiste no principal instrumento utilizado para registro, acompanhamento e controle da execução orçamentária, financeira e patrimonial do Governo Federal (Secretaria do Tesouro Nacional, 2019).

A Tabela 2 descreve, ao final do processo de extração, os atributos coletados que fazem parte deste estudo, com sua codificação, fonte de origem, formato e breve descrição do que é representado.

Tabela 2 - Identificação dos atributos obtidos

Código Atributo	Fonte Origem	Formato	Descrição
ISEO	F1	Categórica (11 classes)	Localização da instalação relativa aos principais setores da estrutura organizacional da MB (apêndice III).
ISNH	F1	Categórica (07 classes)	Identificação do Nível Hierárquico da instalação na estrutura organizacional da MB. Medido pela quantidade de instalações existentes em níveis acima da instalação de referência.
ISTP	F1	Categórica (09 classes)	Categorização do tipo de instalação com base na atividade fim exercida pela mesma. Exemplos: Operativa, Ensino, Saúde, Pesquisa, Diretoria Especializada, etc.
ISPS	F1	Categórica (02 classes)	Indicador de Prestador de Serviço. Identifica uma categoria especial de instalações que não recebem orçamento diretamente e necessitam prestar serviços para outras instalações para se sustentar.
ISPP	F2	Numérica Discreta	Quantidade de pessoas associadas à instalação de referência. Medida composta pela soma de militares e servidores civis.
ESRG	F4	Categórica (05 classes)	Região do País onde se localiza a instalação.
ESDP	F4	Numérica Contínua	Densidade populacional do estado onde se localiza a instalação.
ESIDH	F4	Numérica Contínua	Medida do Índice de Desenvolvimento Humano Municipal (IDH-M) do estado onde se localiza a instalação.
ESRPP	F4	Numérica Contínua	Renda mensal per capita do estado onde se localiza a instalação.
CDDP	F4	Numérica Contínua	Densidade populacional da cidade onde se localiza a instalação.
CLPR	F5	Numérica Contínua	Média mensal de precipitação registrada na estação meteorológica mais próxima da instalação de referência.
CLMAX	F5	Numérica Contínua	Média mensal da temperatura máxima registrada na estação meteorológica mais próxima da instalação de referência.
CLMIN	F5	Numérica Contínua	Média mensal da temperatura mínima registrada na estação meteorológica mais próxima da instalação de referência.

CLHMD	F5	Numérica Contínua	Média mensal da humidade relativa registrada na estação meteorológica mais próxima da instalação de referência.
FNPT	F6	Numérica Contínua	Patrimônio da instalação de referência. Consiste em mensuração financeira de todos os seus bens móveis e imóveis, como computadores, máquinas, móveis, veículos, etc.
FNMQ	F3	Numérica Contínua	Consiste na medição, em metros quadrados, da área construída na instalação de referência.
FNCD	F6	Catagórica (37 classes)	Consiste na classificação da despesa efetuada pela instalação, com base nas categorias presentes nos apêndices I e II.
FNCT	F6	Numérica Contínua	Custo associado a instalação. Consiste em uma fração do gasto que representa aquilo que foi efetivamente consumido. A título de exemplo, gastos para formação de estoques não são contabilizados.

3.4.2 Fase 1 – Transformação

A etapa de transformação demonstrou ser uma das mais complexas desta fase, já que a mesma consiste em integrar os dados obtidos na etapa anterior, de forma a possibilitar sua análise sob diversas perspectivas, além de remover partes irrelevantes ou incompletas que possam comprometer a integridade do estudo. Por se originarem de fontes distintas, os dados foram recebidos em diferentes formatos e em diferentes níveis de detalhamento, fato este que compromete a sua utilização já que se torna obrigatório que o atributo ao menos descreva características que possam ser relacionadas a cada instalação da MB, de forma individualizada. Como resultado desta etapa, que nem todos os atributos coletados puderam ser efetivamente utilizados na Fase 2. A Tabela 3 descreve os principais procedimentos que foram tomados nesta etapa, por fonte de dados:

Tabela 3 - Procedimentos efetuados na etapa de transformação dos dados

Código Fonte Externa de Dados	Descrição da Transformação
F1	Dados foram coletados por acesso direto ao banco de dados de origem, onde estavam armazenados de forma individualizada, por instalação. Cada instalação possuía seu código identificador de forma clara e objetiva. Nenhuma transformação foi necessária.
F2	Foi necessário identificar e remover instalações que estão fora do escopo deste estudo, como por exemplo, instalações localizadas fora do território nacional. Foi necessário agregar a quantidade de pessoas por instalação para remover informações ligadas ao posto/graduação de cada militar, irrelevantes para este estudo.
F3	Dos dados obtidos, foi possível relacionar cada instalação com o seu Estado/Município de

Origem.

Entretanto, a informação mais relevante desta fonte, que seria a área construída de cada instalação, medida em metros quadrados, não pode ser utilizada neste estudo. Uma análise dos dados identificou que a fonte externa não armazena os registros por instalação, e sim por grandes aglomerados de instalações em uma determinada localização geográfica. Dentre os diversos aglomerados, o maior deles comportava mais de 48 instalações, sendo impossível identificar a medida de forma individualizada para integrá-la aos demais atributos.

F4 Foi necessário criar uma codificação que pudesse relacionar as cidades e estados brasileiros oriundos desta fonte com a relação gerada na fonte F2. Posteriormente, passou-se a remover dados considerados irrelevantes para este estudo, como por exemplo, nome do Governador do estado, receitas realizadas e total de veículos registrados. Além disso, no caso dos municípios, os dados foram obtidos em 27 planilhas que, além do procedimento de remoção já citado, necessitaram passar por um processo de agregação e padronização, de forma a se relacionarem corretamente com seus estados de origem.

F5 No período da coleta dos dados (junho de 2019), a fonte externa não permitia uma exportação dos dados de todas as estações meteorológicas do país. Portanto, foi necessário coletar os dados sobre cada estação, individualmente. Outro complicador foi o fato de não existirem estações meteorológicas em alguns municípios onde estão localizadas instalações da MB. Para contornar este problema, a estação meteorológica mais próxima foi utilizada, desde que a mesma se mantivesse, aproximadamente, na mesma altitude (medida que influencia condições climáticas). Além disso algumas estações meteorológicas não fornecem todos os dados utilizados neste estudo e, portanto, foram tratados como valores ausentes.

F6 Por se tratar de um sistema contábil do governo federal, os valores relacionados aos gastos são rigorosamente associados a cada instalação, o que facilitou a integração dessa fonte externa. Foi necessário, entretanto, remover registros que representam recursos da MB utilizados em instalações não pertencentes à organização. Entretanto, para os valores relacionados ao patrimônio de cada instalação foi identificada a presença de aglomerados semelhantes aos encontrados para o atributo FNMQ em F3, sendo impossível identificar a medida de forma individualizada para integrá-la aos demais atributos.

Por fim, é importante ressaltar que o fim desta etapa representa os esforços necessários para que se possa atingir o OE1 “Identificar, coletar e analisar os dados atualmente disponibilizados que representem atributos das instalações da organização e que possam auxiliar no desenvolvimento do modelo.”. Entretanto, transformações adicionais são efetuadas na fase 2, dessa vez com o foco no treinamento de modelos de ML.

3.4.3 Fase 1 – Carga

A etapa de carga se iniciou com as planilhas efetivamente tratadas da etapa anterior, que foram utilizadas como base para o desenvolvimento de uma infraestrutura capaz de receber e gerenciar os dados. Esta etapa guarda estrita relação com o OE2 “Desenvolver uma infraestrutura capaz de gerenciar os dados coletados de forma eficaz, possibilitando o relacionamento entre os mesmos, bem como diversos tratamentos e abordagens.”

Inicialmente, foi confeccionado um Diagrama Entidade-Relacionamento (DER), idealizado por Chen (1976), e muito utilizado até os dias atuais como um modelo para descrever os dados ou aspectos dos mesmos dentro de um domínio específico e para suportar a arquitetura de dados de uma organização, orientando a implementação de um banco de dados. A Figura 8 apresenta o DER, e ilustra como os dados foram organizados para favorecer a consulta e a análise. Cabe ressaltar que a arquitetura de dados foi projetada de maneira a facilitar a inserção de recursos adicionais no futuro.

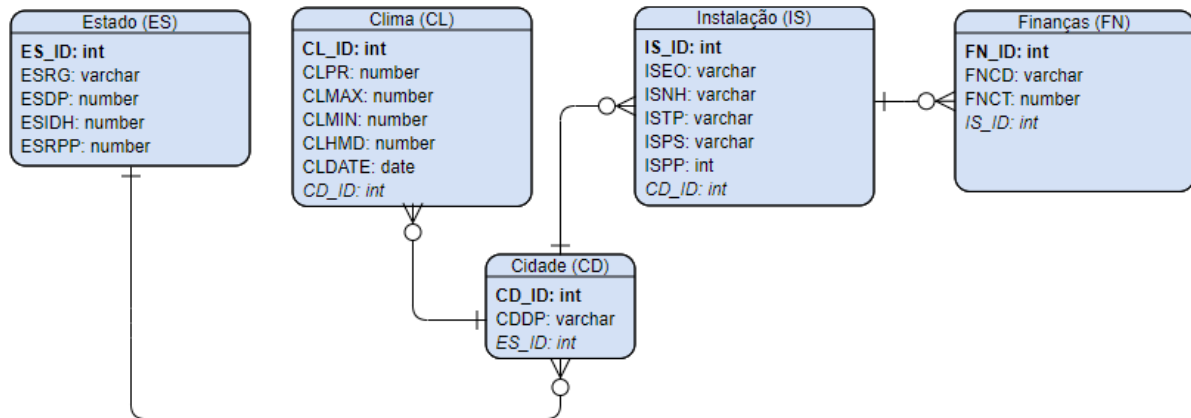


Figura 8 - Diagrama Entidade-Relacionamento

Após o DER, os dados foram carregados em um Banco de Dados MySQL⁴ que, por meio de consultas, foi capaz de gerenciar os dados e apresentá-los sob diversas perspectivas que serão utilizadas na fase 2.

3.4.4 Fase 2 – Entrada de Dados

Esta etapa representa o início da Fase 2, onde o algoritmo já mencionado é executado com o propósito de receber os dados da Fase 1 e aplicar uma série de técnicas baseadas em ML para preparar e analisar os dados, treinar os modelos e avaliar os resultados.

Inicialmente, o algoritmo carrega um arquivo SQL⁵ contendo as instruções para efetuar uma consulta ao Banco de Dados MySQL que, ao ser efetuada, retorna os dados solicitados, com a devida identificação de quais são os Atributos-Alvo (AA) a serem estudados. Desta forma, ao construir cada consulta previamente, o algoritmo pode ser carregado com diferentes perspectivas sobre os dados, a cada execução. Para este estudo foram idealizadas duas perspectivas, daqui em diante chamadas de

⁴ MySQL é um sistema de gerenciamento de banco de dados relacionais, de código aberto, gerenciado pela empresa Oracle. A versão utilizada neste estudo (*Community Edition*) foi retirada de <https://www.mysql.com/products/community/>

⁵ *Structured Query Language*, uma das linguagens mais utilizadas para manipulação de dados em bancos de dados

P1 e P2, que permitem fornecer ao algoritmo os dados coletados na Fase 1 em diferentes estruturas, representando diferentes AA a serem analisados pelo processo de aprendizagem.

A variável categórica FNCD, responsável pela classificação da despesa efetuada com base em sua natureza, foi utilizada como base para a criação das perspectivas, aliado ao entendimento da organização de quais despesas pertencem a categoria FC e MN (apêndices I e II). A Figura 9 ilustra como as categorias da variável FNCD se relacionam com P1 e P2.

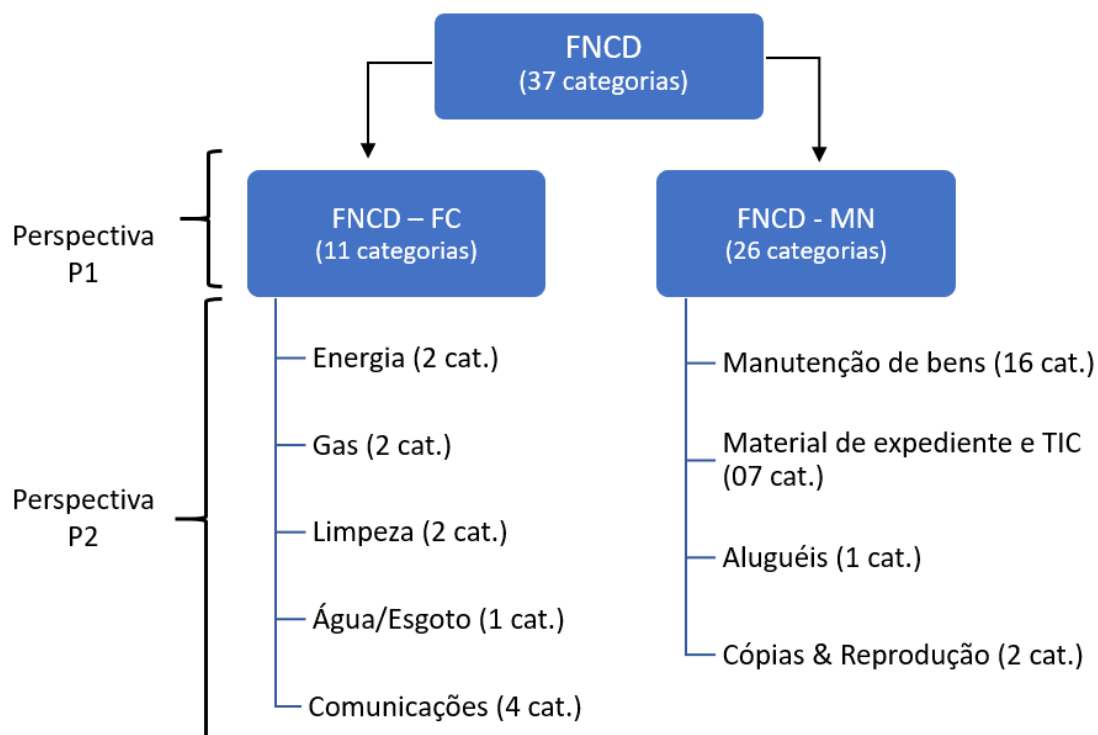


Figura 9 - Relacionamento entre as categorias de FNCD com a criação de perspectivas.

Em resumo, enquanto o objetivo em P1 é ter dois modelos finais, um focando todos os gastos em FC agregados e outro capaz de explicar todos os gastos em MN, em P2 o algoritmo treinará um total de nove modelos: Energia (ENE) , Gás (GAS), Limpeza (LIMP), Água / Esgoto (AGEG) e Comunicações (COM) para os gastos em FC e Manutenção de Bens (MNT), Material de Expediente e TIC (XPTI), Aluguéis (ALG) e Cópias e Reprodução (CPRP) para os gastos em MN. Uma das principais razões para a existência de P1 e P2 é verificar se os modelos finais terão performance superior ao se especializarem em uma categoria específica (limpeza, por exemplo), ou se o processo de aprendizagem é capaz de lidar com diferentes categorias de despesa em um só modelo. A Tabela 4 resume a definição dos AA a serem analisados de acordo com as perspectivas desenvolvidas.

Tabela 4 - Descrição dos Atributos-Alvo (AA) definidos

Sigla do AA	Descrição
P1 FC	Atributo-Alvo: Soma de todas as despesas do tipo FC (Energia, Gás, Limpeza, Água/Esgoto e Comunicações)
P2 GAS	Atributo-Alvo: Despesas de Gás
P2 LIMP	Atributo-Alvo: Despesas de Limpeza
P2 ENE	Atributo-Alvo: Despesas de Energia Elétrica
P2 AGEG	Atributo-Alvo: Despesas de Água e Esgoto
P2 COM	Atributo-Alvo: Despesas de Comunicações
P1 MN	Atributo-Alvo: Soma de todas as despesas do tipo MN (Manutenção de Bens, Material de Expediente/TIC, Aluguéis e Cópias/Reprodução)
P2 MNT	Atributo-Alvo: Despesas de Manutenção de Bens
P2 XPTI	Atributo-Alvo: Despesas Material de Expediente/TIC
P2 ALG	Atributo-Alvo: Despesas de Aluguéis
P2 CPRP	Atributo-Alvo: Despesas de Cópias/Reprodução

3.4.5 Fase 2 – Atributos Categóricos

Os atributos categóricos contêm valores não numéricos, e nem sempre podem ser utilizados diretamente em um processo de aprendizado de ML. Para contornar este problema, os mesmos necessitam de tratamento especial capaz de transformá-los em um formato numérico adequado, de forma a que contribuam para a explicação do atributo alvo.

Nesta etapa, todos os atributos com valores de texto que representam categorias são identificados pelo algoritmo para que sejam, por ocasião do treinamento de modelos, transformados pela técnica *one-hot encoding*, que transforma um atributo categórico de x classes e y observações em x atributos binários com y observações cada. A Figura 10 ilustra a aplicação da técnica mencionada em uma variável categórica fictícia.

COR
verde
azul
vermelho
azul
verde

COR.vermelho	COR.verde	COR.azul
0	1	0
0	0	1
1	0	0
0	0	1
0	1	0

Figura 10 - Exemplo de one-hot encoding em variável categórica representando cores

3.4.6 Fase 2 – Valores Ausentes

As funcionalidades da linguagem R podem ser estendidas por meio de pacotes. Na etapa responsável pelo tratamento de valores ausentes nos dados coletados, o pacote “mice”, cuja técnica é descrita por van Buuren e Groothuis-Oudshoorn (2011), foi utilizado para prover um modelo de

imputação múltipla de dados cujo propósito é, principalmente, preservar as relações existentes entre os atributos e também a incerteza entre os mesmos. O pacote também possui funcionalidades voltadas para a análise dos valores ausentes e também dados imputados como resultado.

A Figura 11 apresenta um dos resultados da análise de valores ausentes gerados pelo pacote “mice”, durante a execução da perspectiva P1 – FC. A título de exemplo de sua interpretação, a segunda linha da figura informa que existem 43 registros com valores ausentes em exatamente 4 atributos (CLPR, CLMAX, CLMIN e CLHMD). A Figura 11 apresenta um gráfico onde é possível observar os valores numéricos imputados (pontos em vermelho) relativamente aos dados originais (pontos em azul).

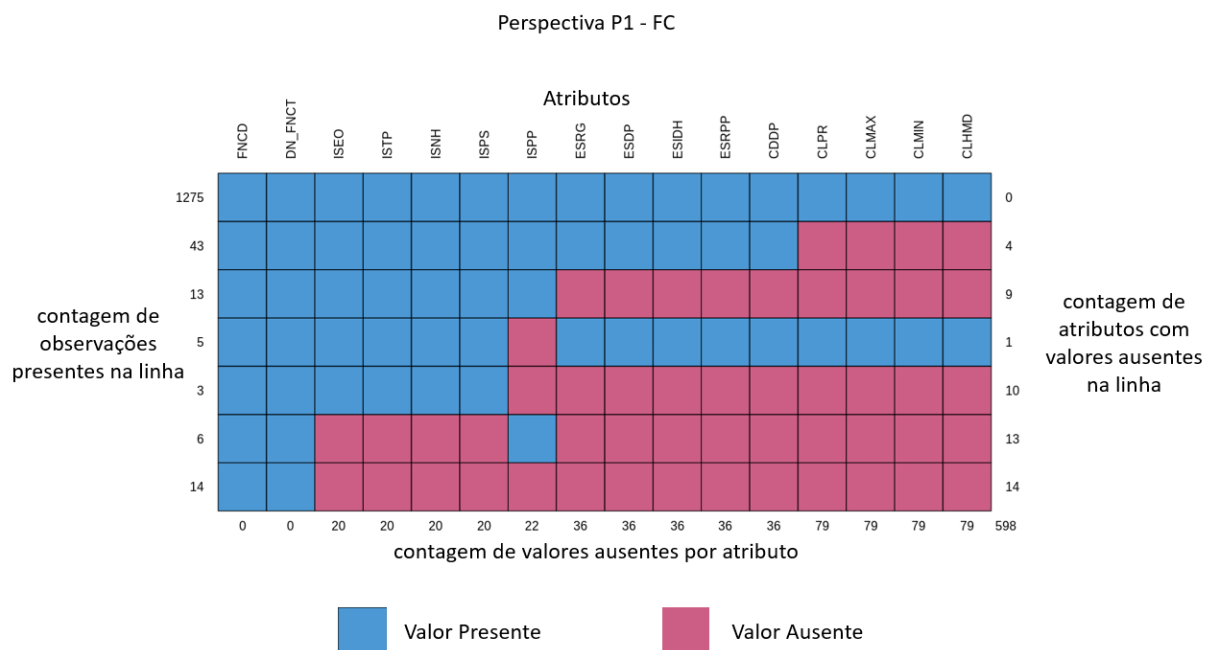


Figura 11 - Análise dos valores ausentes em P1-FC realizado pelo pacote “mice”

O resultado desta etapa, que efetua a imputação somente em atributos numéricos, se propõe a fornecer uma maior quantidade de observações válidas para o processo de aprendizado. Estudos mostram que uma abordagem voltada para a simples remoção de registros com qualquer valor ausente pode resultar na perda de grande parte dos dados coletados, além de introduzir distorções nos relacionamentos entre os atributos (Myrtveit, Stensrud, & Olsson, 2001). Entretanto, aquelas observações que possuem atributos categóricos com valores ausentes são removidas.

3.4.7 Fase 2 – Seleção de Atributo Alvo (AA)

Esta etapa foi incluída no algoritmo com o propósito de torná-lo capaz de receber dados com mais de um AA para ser estudado. Neste caso, as etapas 2.5 a 2.9 da Figura 12 serão repetidas para cada AA identificado nos dados recebidos, formando assim um ciclo que é encerrado após todas terem sido analisadas e seus modelos devidamente avaliados.

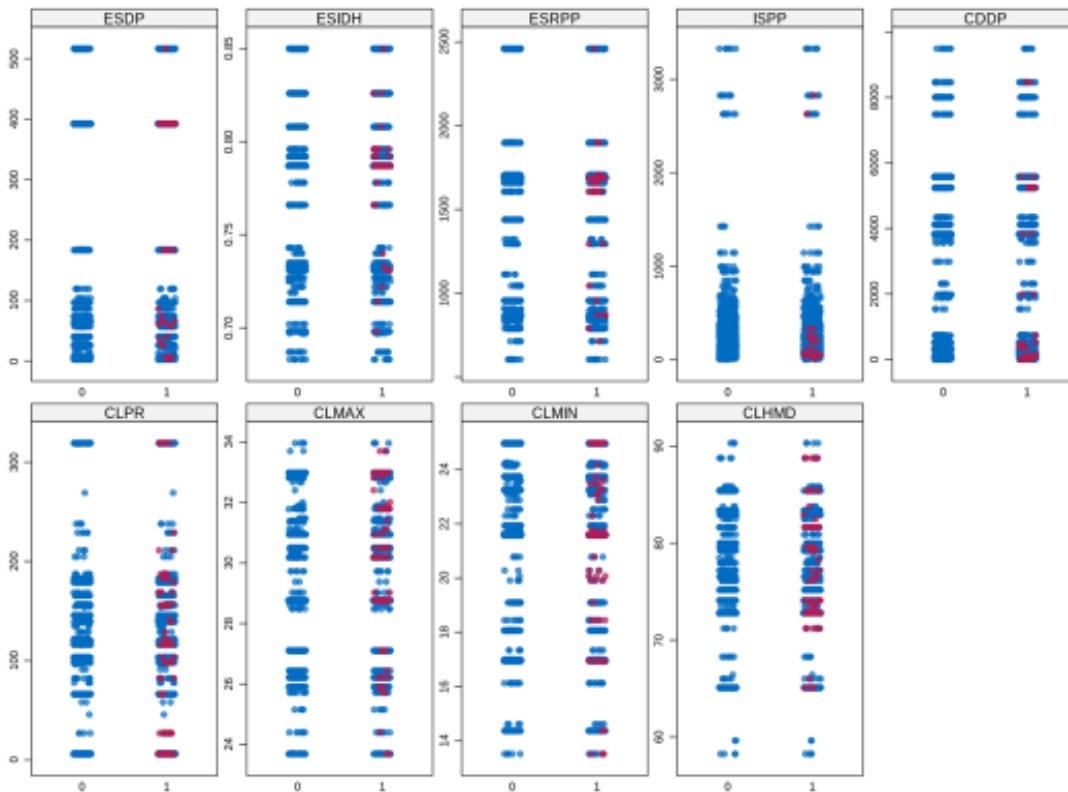


Figura 12 - Análise dos valores imputados em P1-FC

3.4.8 Fase 2 – Particionamento de Dados

Conforme estudado por Hawkins (2004), diversos métodos baseados em medições quantitativas para ajustar modelos matemáticos e estatísticos para solução de problemas sofrem com a possibilidade de *overfitting* ou sobre ajuste (tradução minha), que ocorre quando um modelo utiliza abordagens mais complexas do que as necessárias para ter uma boa performance. O autor também aponta que a origem do *overfit* pode estar em diversos fatores, sendo um dos principais a escolha do próprio modelo, por exemplo: utilizar redes neurais, conhecido por ser um modelo flexível capaz de reconhecer padrões não lineares, para se adaptar a dados cujo relacionamento obedece em sua maior parte o modelo linear poderá resultar em uma abordagem demasiadamente complexa que reconhecerá

padrões que são irrelevantes ou que podem até piorar o resultado final. De forma análoga, o oposto também pode ocorrer e é definido como *underfit* ou sub ajuste (tradução minha) que ocorre quando um modelo não é capaz de aprender relacionamentos importantes para a performance final do modelo. A Figura 13 ilustra o comportamento de três modelos fictícios onde o primeiro utilizou uma abordagem linear, ignorando relacionamentos relevantes, e o terceiro foi flexível o suficiente para praticamente memorizar os dados ao invés de generalizá-los. Como resultado, espera-se que ambos os modelos tenham performances ruins quando utilizados em previsões com base em dados novos, não utilizados no treinamento.

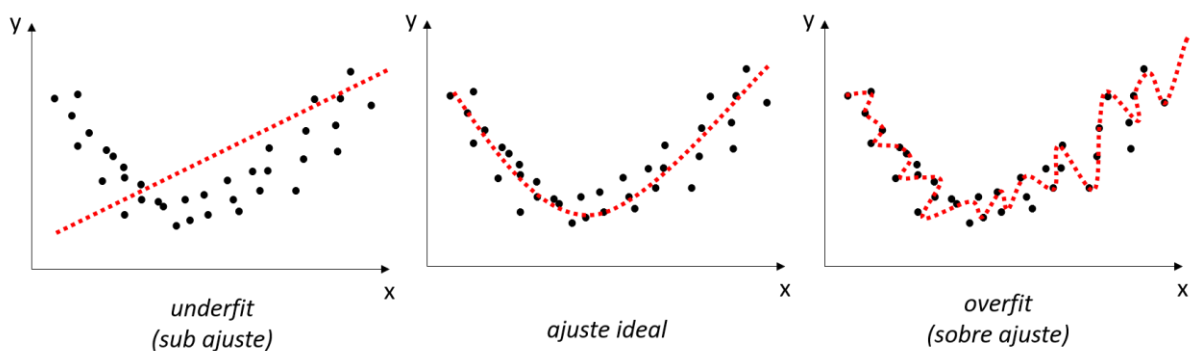


Figura 13 - Exemplo gráfico de underfit e overfit no ajuste de modelos

Os problemas expostos acima reforçam a necessidade de se validar o modelo, de forma a entender a sua capacidade em termos de *overfitting* e *underfitting* (Bilal & Oyedele, 2020). Essa validação consiste em solicitar que o modelo efetue previsões cujo valores já são conhecidos, de forma a avaliar sua precisão. Uma das técnicas mais utilizadas para este fim é *k-folds cross-validation*, ou “validação cruzada em k dobras” (tradução minha), que consiste em dividir os dados em k subconjuntos, treinar o modelo com $k-1$ subconjuntos e testar a performance do mesmo com o subconjunto restante. O procedimento deve ser repetido k vezes de forma que cada subconjunto possa ser utilizado para validação; ao final, a performance de todos os modelos gerados são agregadas (geralmente pela média e desvio padrão e de acordo com a métrica utilizada) para que se obtenha uma medição final da performance (Kuhn & Johnson, 2013). Diversos estudos já comprovaram sua eficácia para validação de modelos (Hawkins, 2004; Kohavi, 1995) e sugerem, quando possível, um tamanho de ao menos 50 observações em cada subconjunto para obtenção de resultados confiáveis (Hawkins, 2004). A Figura 14 apresenta uma representação gráfica conceitual do método.

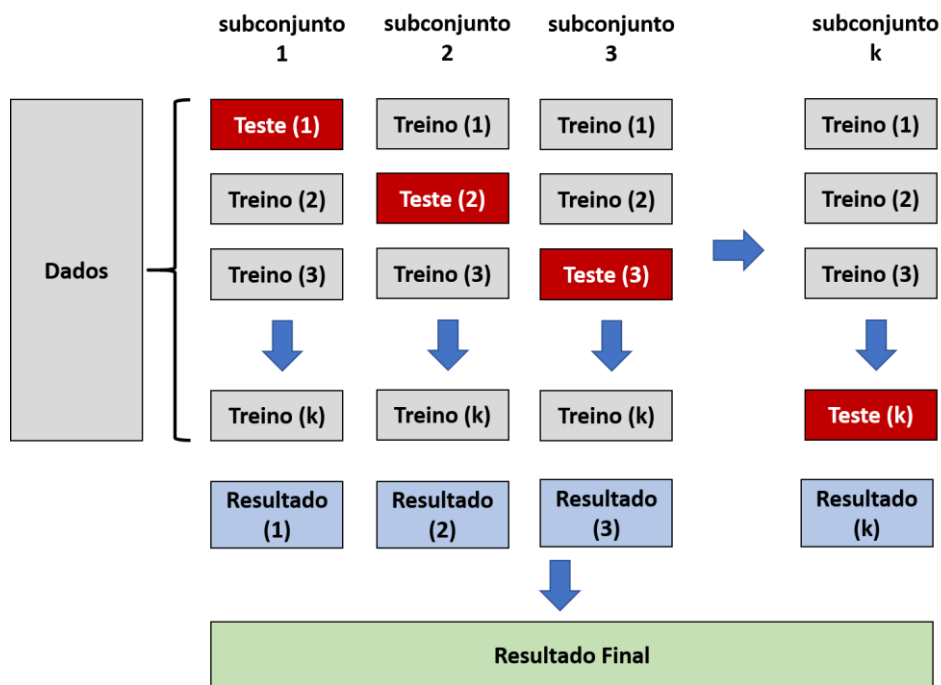


Figura 14 - Representação gráfica do método k-folds cross-validation

Para este estudo, o método foi utilizado com uma quantidade k igual a 10, ou seja, com a criação de 10 subconjuntos para aplicação dos procedimentos mencionados. Entretanto, caso a quantidade de observações em cada subconjunto seja menor que 50, a quantidade de subconjuntos a serem formados é reduzida até que a restrição seja obedecida. Todo o método é, então, integralmente repetido 5 vezes, conforme uma ligeira variação do método descrita por Kuhn e Johnson (2013).

3.4.9 Fase 2 – Seleção de Atributos

A seleção prévia dos atributos mais importantes a serem utilizados no momento do treinamento dos modelos é uma etapa essencial para obter bons resultados, guardando estrita relação com as possibilidades de *overfit* e *underfit*. De acordo com Hawkins (2004) e Kohavi e John (1997), modelos que incluem atributos desnecessários são naturalmente mais complexos e levam a decisões piores por considerar variações que, para o problema específico, tem comportamento aleatório.

Para cumprir esta tarefa do algoritmo, foi utilizado o pacote “Boruta”, descrito por Kursa e Rudnicki (2010), que se propõe a identificar os atributos mais relevantes para explicar a variabilidade do AA, oferecendo, portanto, uma forma prática de reduzir a complexidade do modelo final sem impactar sua performance preditiva. Seu funcionamento é baseado no desenvolvimento de um modelo temporário de classificação que identifica aqueles atributos que se provaram, estatisticamente, ser menos relevantes do que atributos de variabilidade aleatória. A Figura 15 contém um dos resultados da seleção de atributos fornecida ao analisar a perspectiva P1 – FC, onde pode-se observar a importância

de cada atributo quando comparadas com dados aleatórios. Os atributos identificados como irrelevantes ou duvidosos são removidos das futuras etapas do algoritmo.

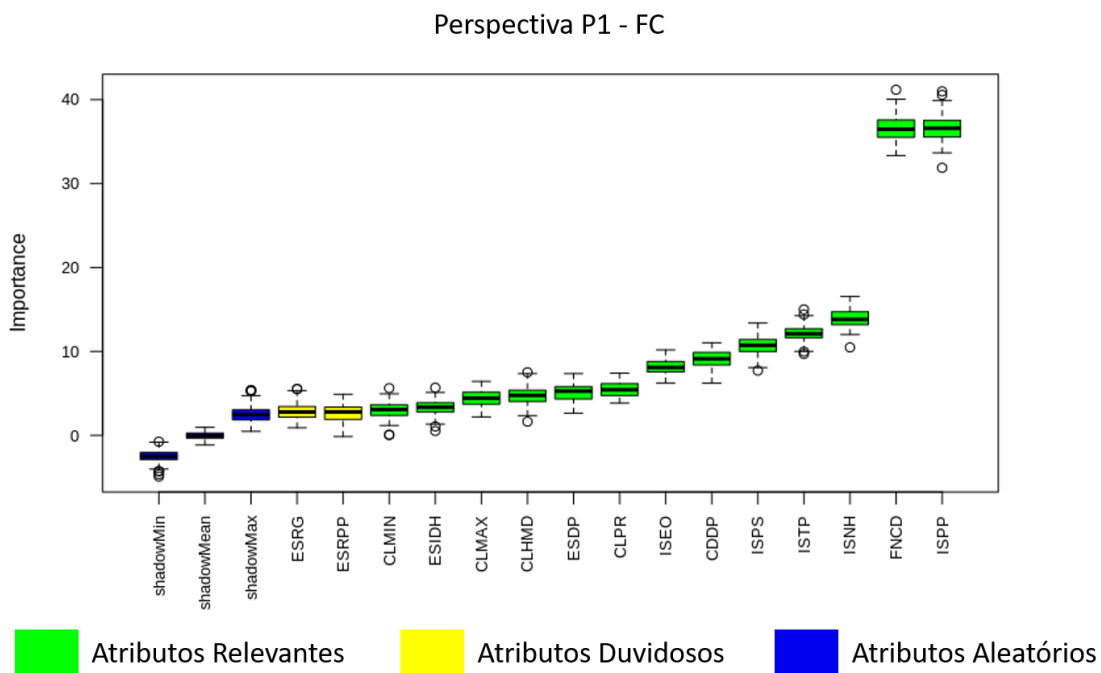


Figura 15 - Análise da seleção de atributos realizado pelo pacote "Boruta"

3.4.10 Fase 2 – Identificação de *Outliers*

Conforme estudado por Hodge e Austin (2004), a identificação de *Outliers* assume um papel crítico em muitos contextos por ser capaz de indicar condições anormais que podem resultar em significativa degradação de performance. O autor também lista diversas aplicações que utilizam esta abordagem, como detecção de fraudes e intrusos, monitoramento de atividades e performance, diagnóstico de falhas, entre outras.

Definido como “uma observação (ou subconjunto de observações) que aparentam ser inconsistentes quando comparadas com o restante do conjunto de dados” (Barnett & Lewis, 1978, p. 5), o julgamento de *outliers* por parte do observador deve ser realizado com sensatez, já que o objetivo principal é identificar se essas observações, que aparentam ser inconsistentes, são ou não membros efetivos da população que se analisa. Caso não sejam, as inferências que sejam feitas sobre a população podem sofrer distorções (Barnett & Lewis, 1978).

Os dados utilizados neste estudo representam características das instalações da MB e seus respectivos gastos em FC e MN. Durante o processo de coleta pelo pesquisador, os dados foram analisados e filtrados por possíveis erros e inconsistências. Além disso, por características próprias da Organização em foco, é esperado que os dados apontem a existência de *Outliers*, devido a existência as algumas poucas instalações de grande tamanho que centralizam atividades específicas que não são realizadas por nenhuma outra instalação da MB.

Desta forma, esta etapa foi implementada somente para fins didáticos, pois pode gerar informações úteis ao identificar *Outliers* em diferentes perspectivas. Entretanto, nenhum tratamento adicional é realizado após a identificação, já que se tratam de membros efetivos da população que se estuda. O pacote “solitude”, descrito por Liu, Ting, e Zhou (2008), explora a característica básica dos *Outliers* de serem “poucos e diferentes do restante” para efetuar uma abordagem baseada em um algoritmo de ML chamado *árvores de decisão*, capaz de isolar observações anormais.

3.4.11 Fase 2 – Treinamento de Modelos: Pré-Processamento

Previamente ao efetivo treinamento de modelos, é necessário garantir que os dados que serão utilizados no processo de aprendizado estão em formato adequado, o que pode vir a causar uma série de transformações adicionais, além daquelas citadas na seção 3.4.2. Zhang, Zhang, e Yang (2003) apontam que o pré-processamento é uma fase fundamental na aplicação de ML, que pode vir a causar significativos impactos na performance final dos modelos. O pacote “caret” (Kuhn, 2008) foi utilizado por ser capaz de padronizar o procedimento a ser adotado para aplicação de algumas das técnicas de pré-processamento a seguir:

- *Centering e Scaling*, ou Centralizar e Ajuste de Escala (tradução minha) : Kuhn & Johnson (2013) referem-se como a transformação de dados mais direta e comum a ser realizada. Com a Centralização, o valor relativo à média é subtraído de todas as observações, resultando em um atributo com média zero. Já no Ajuste de Escala, cada valor é dividido pelo desvio padrão das observações, resultando em um atributo com desvio padrão de valor igual a 1. Assim, o efeito de atributos com diferentes unidades de medida e, portanto, diferentes grandezas numéricas é anulado (Huang, Li, & Xie, 2015).
- *Box-Cox*: Idealizado por Box e Cox (1964), consiste no cálculo e análise de um parâmetro (λ) para selecionar, dentre um grupo de transformações, a mais adequada para aprimorar assimetrias na distribuição dos dados em foco. A seguinte equação explica a análise do parâmetro λ , onde x é o valor da observação, e x^* é o valor final transformado:

$$x^* = \begin{cases} \frac{x^\lambda - 1}{\lambda} & \text{se } \lambda \neq 0 \\ \log x & \text{se } \lambda = 0 \end{cases}$$

A título de exemplo, uma breve análise da distribuição do atributo ISPP (quantidade de pessoas) foi efetuada durante a perspectiva P1-FC, o que resultou $\lambda = 0$. A Figura 16 apresenta que, apesar do resultado da transformação não ser exatamente simétrico, a mesma passou a guardar certa similaridade com a distribuição normal. Alguns modelos de ML possuem melhor performance quando lidam com dados que se aproximam da distribuição log-normal (Ghalekhondabi et al., 2017). Cabe, portanto, ressaltar que esta análise é efetuada para cada atributo em cada execução do algoritmo, uma vez que qualquer modificação na forma de se obter os dados pode alterar sua distribuição e

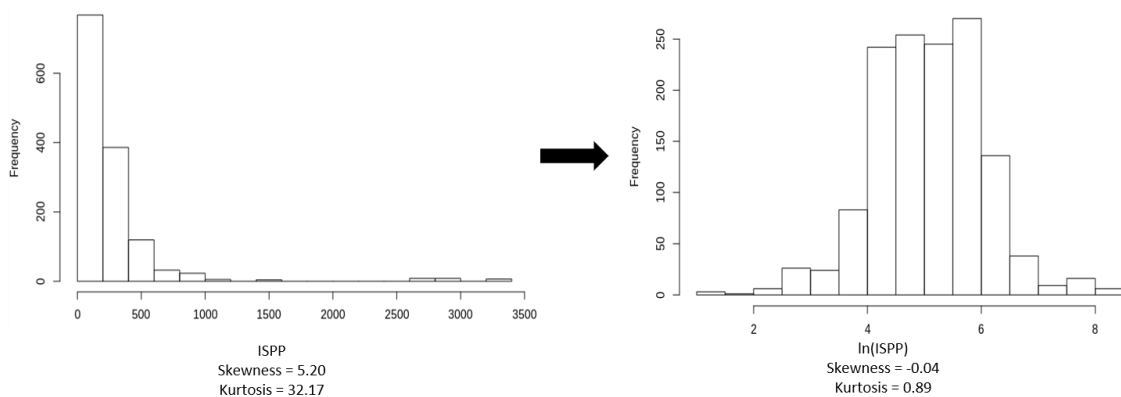


Figura 16 - Histograma do atributo ISPP antes e depois da transformação.

requerer uma nova avaliação.

3.4.12 Fase 2 – Treinamento de Modelos: Escolha e Configuração

Após o pré-processamento, cada iteração do algoritmo treina 5 modelos de ML para regressão, onde os atributos são utilizados para obter previsões sobre o AA. A escolha dos modelos a serem treinados é uma etapa fundamental para a performance final das previsões que se desejam obter, já que cada modelo possui suas vantagens e desvantagens (Hawkins, 2004; Kuhn & Johnson, 2013; Lee et al., 2019). O pacote “caret” (Kuhn, 2008) foi utilizado por oferecer meios eficazes para padronização no treinamento de modelos distintos, bem como por ser capaz de repetir o processo com diversas combinações dos parâmetros iniciais de cada modelo, selecionando aquela combinação que resultou nos melhores resultados.

Para fins de estudo, e também para possibilitar que esta abordagem conceitual seja aplicada em outros problemas de interesse da organização, esta etapa não se reduzirá a escolha de um único modelo, mas sim no treinamento de modelos diversos, para fins de comparação. Os seguintes modelos são treinados:

- **Regressão Linear Múltipla (RLM):** Trata-se do modelo mais simples de ML, e também uma técnica largamente utilizada por outras abordagens não relacionadas a ML. Consiste em encontrar um plano que solucione a seguinte equação:

$$y = \beta_0 + \beta_1x_1 + \beta_2x_2 + \dots + \beta_nx_n$$

Onde, y é o atributo-alvo (AA), $x_{1\dots n}$ são os atributos utilizados (de um total de n atributos) e $\beta_{0\dots n}$ são pesos (coeficientes) associados a cada um deles. Quando o plano não é suficiente para interligar todas as observações, o plano é ajustado de tal forma a minimizar o erro, mensurado pela distância entre o valor previsto e o verdadeiro.

Como principal vantagem, sua simplicidade oferece alta interpretabilidade dos resultados por meio da avaliação dos coeficientes associados a cada atributo, que podem ser entendidos como um grau de importância do mesmo para explicar o AA. Entretanto, o modelo assume que o relacionamento entre os atributos é linear e que os mesmos obedecem a distribuição normal, além de ser bastante sensível a presença de *outliers* (Kuhn & Johnson, 2013). Sua aplicação neste estudo tem o propósito de oferecer uma linha de base para fins de comparação com os outros modelos, e também para apontar relações lineares não esperadas.

- **Support Vector Regression (SVR):** Trata-se de um modelo extremamente flexível e eficaz para a realização de previsões, sendo facilmente adaptável para problemas não-lineares e significativamente robusto contra a influência de *outliers* (Kuhn & Johnson, 2013). Consiste em uma abordagem cujo objetivo é buscar a generalização, ou seja, minimizar a probabilidade com a qual o modelo treinado produzirá erros ao lidar com novos dados. Seu mecanismo é significativamente mais complexo quando comparado com RLM, o que também compromete a interpretação de seus resultados. Entretanto o estudo efetuado por Smola e Schölkopf (2004) apresenta um detalhamento de seu funcionamento, incluindo uma explicação sobre seus *kernels* - funções adicionais, escolhidas pelo operador, para generalizar o modelo e englobar funções não-lineares (polinomiais, radiais, hiperbólicas, etc.). Os autores também sugerem a utilização de uma *kernel* do tipo gaussiana quando pouco se sabe sobre a estrutura dos dados que serão analisados, sendo este tipo específico

de *kernel* amplamente utilizado em outras pesquisas. Hong et al. (2017), por exemplo, realizaram uma comparação de *kernels* para mapear suscetibilidades a deslizamentos de terra com base em dados geográficos onde a do tipo gaussiana obteve a melhor performance.

- *Random Forest* (RF): Modelo baseado no conceito de árvores de decisão, que consiste em um método simples e eficiente de regressão, onde uma ou mais declarações condicionais do tipo “se-então” dividem os dados em regiões sem sobreposição que são utilizadas para efetuar as previsões (Kuhn & Johnson, 2013; Quinlan, 1986). A Figura 17 ilustra, à esquerda, uma simples árvore de decisão com duas divisões e três terminais (ou folhas), onde X_1 e X_2 são atributos e Y é o AA e, à direita, como as regiões são representadas nos dados para realização das previsões.

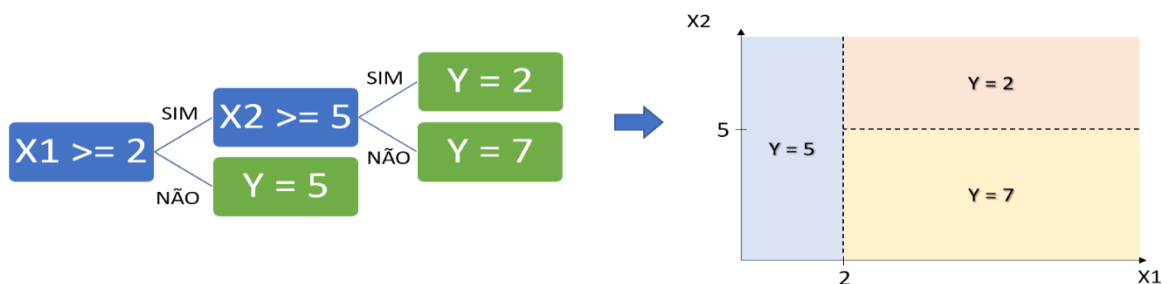


Figura 17 - Representação de uma Árvore de Decisão e suas consequências nos dados

Dessa forma, o método *Random Forest* foi idealizado por Breiman (2001) para lidar com algumas das diversas limitações das árvores de decisão como baixa performance preditiva e instabilidade em caso de pequenas alterações nos dados. Trata-se de um conjunto de árvores de decisão (a quantidade exata é um parâmetro configurável deste método), criadas de forma independente, que utilizam em sua criação atributos selecionados de forma completamente aleatória. A componente aleatória adicionada ao método o torna extremamente robusto contra *outliers* e rápido em termos computacionais, já que as árvores são independentes e podem ser criadas em paralelo, mas que podem sofrer de *overfit* em alguns casos (Chambers, Hentges, & Zhao, 2004; Hodge & Austin, 2004; Kuhn & Johnson, 2013).

- *Gradient Boosting Machines* (GBM): Método idealizado por Friedman (2001), que também utiliza árvores de decisão como base, mas é implementado de forma significativamente distinta: uma árvore de decisão inicial é criada e seus erros são mensurados; uma nova árvore de decisão é criada sobre os erros da anterior, em uma tentativa de minimizá-los, sendo posteriormente integrado ao modelo inicial. O procedimento se repete até uma

quantidade máxima de iterações (um dos parâmetros do método), de forma que, a cada repetição uma nova árvore de decisão é desenvolvida para aprimorar as regiões com pior performance (Touzani, Granderson, & Fernandes, 2018). Diferentemente do modelo RF, as árvores de decisão são criadas de forma sequencial, e, por isso, existe a tendência de ser um método mais lento em termos computacionais.

- *XGBoost* (XGB): Modelo que também se baseia na técnica de “boosting” do modelo GBM, mas que conta com diversas características adicionais, como processamento em paralelo, abordagem contra valores ausentes, regras de penalização – uma técnica que penaliza abordagens complexas para evitar *overfit*, entre outras. Tais fatores o tornam computacionalmente eficiente na resolução dos mais diversos tipos de problemas, tendo chamado a atenção de pesquisadores da área nos últimos anos ao ser utilizado de forma consistente em diversas competições de ML para fornecer os melhores resultados no menor tempo (T. Chen & Guestrin, 2016). É possível encontrar diversas pesquisas onde este modelo obteve a melhor performance, como por exemplo no estudo de Robinson et al. (2017)

3.4.13 Fase 2 – H2O *Automatic Machine Learning*

Em paralelo com as etapas 2.5 a 2.8 da Figura 7 nesta etapa é realizada uma abordagem diferente ao se utilizar uma ferramenta de *Automatic Machine Learning* (AutoML). Plataformas de aprendizado automático é um campo de estudo recente que visa criar procedimentos automatizados e genéricos que possam ser aplicados em quaisquer problemas e estrutura de dados. Lee et al. (2019) explicam que algumas áreas lidam com dados sensíveis e incorrem em altos custos para permitir que especialistas os estudem para gerar conhecimento sem comprometer a segurança dos dados. Dessa forma, essas áreas podem se beneficiar de plataformas que requerem pouca ou nenhuma intervenção humana. Os autores também acreditam ser possível desenvolver uma plataforma completamente automática em um futuro próximo, já que a maioria das funcionalidades a serem utilizadas já podem ser desenvolvidas com a tecnologia atual.

H2O AutoML consiste em uma plataforma que visa simplificar o processo de aprendizado, ao permitir que usuários não especializados possam utilizar e usufruir do potencial do ML. AutoML lida com todas as etapas de pré-processamento e treinamento por conta própria, inclusive ao decidir quais modelos serão treinados e validados, exigindo para tal poucos parâmetros iniciais, como por exemplo, tempo máximo de execução e/ou número máximo de modelos. Dentre os métodos disponíveis para

treinamento estão GBM, XGB, Redes Neurais (*Artificial Neural Networks* – ANN) e *Stacked Ensembles* – uma técnica que permite gerar um modelo final baseado na combinação das previsões de um conjunto de modelos distintos previamente treinados (Boehmke & Greenwell, 2019). Conforme aponta Van der Laan, Polley, e Hubbard (2007), *Stacked Ensembles* tendem a obter uma performance superior a todos os modelos individuais utilizados em seu desenvolvimento, e por isso são considerados o sistema de aprendizado mais próximo do que seria o ótimo.

Neste estudo, o H2O AutoML foi executado com um tempo limite de análise de 15 minutos para cada AA. Os resultados são apresentados por meio da funcionalidade “*leaderboard*”, que fornece um ranking dos modelos treinados e validados durante a execução, com base em diversas métricas. O algoritmo então seleciona o modelo com a melhor performance para as etapas posteriores. A Figura 18 apresenta os resultados obtidos pela ferramenta na perspectiva P1-FC, onde os *Stacked Ensembles* obtiveram melhor performance, seguido de três modelos baseados em GBM e um em XGB.

	model_id	mean_residual_deviance	rmse	mse	mae	rmsle
1	StackedEnsemble_AllModels_AutoML_20200317_111555	2.681229	1.637446	2.681229	1.218639	0.2144365
2	StackedEnsemble_BestOfFamily_AutoML_20200317_111555	2.739878	1.655258	2.739878	1.231559	0.2165692
3	GBM_3_AutoML_20200317_111555	2.780302	1.667424	2.780302	1.241851	0.2179510
4	GBM_grid__1_AutoML_20200317_111555_model_9	2.800637	1.673510	2.800637	1.253527	0.2180331
5	GBM_2_AutoML_20200317_111555	2.804740	1.674736	2.804740	1.249814	0.2181614
6	XGBoost_1_AutoML_20200317_111555	2.808992	1.676005	2.808992	1.258875	0.2176965

Figura 18 - Funcionalidade "leaderboard" do H2O AutoML

3.4.14 Fase 2 – Avaliação dos Resultados

Esta etapa se compromete a definir métricas que permitem avaliar a performance dos modelos criados na etapa anterior. Torna-se crucial estabelecer uma métrica comum entre todos os modelos, já que os mesmos são implementados e calculados de diferentes formas. Os pacotes da linguagem R utilizados durante esta abordagem conceitual, bem como diversas pesquisas encontradas durante a revisão da literatura consolidam a utilização de 3 métricas principais para problemas de regressão:

- *Mean Absolute Error (MAE)*: Consiste na média de todos os erros mensurados pelas diferenças entre os valores que representam as observações reais do AA (y) e as suas respectivas previsões feitas pelo modelo (\hat{y}). Considerando a existência de n observações, a fórmula é a seguinte:

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i|$$

Cabe ressaltar que, na mensuração do erro, o MAE utiliza o conceito de distância e não considera a direção do mesmo (erros positivos ou negativos). Portanto, o erro calculado tem seu sinal removido pela função modular.

- *Root Mean Square Error (RMSE)*: Consiste na raiz quadrada da média dos quadrados dos erros mensurados pelas diferenças entre os valores que representam as observações reais do AA (y) e as suas respectivas previsões feitas pelo modelo (\hat{y}). Considerando a existência de n observações, a fórmula é a seguinte:

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2}$$

Cabe ressaltar que o RMSE também utiliza o conceito de distância, uma vez que o erro calculado é elevado ao quadrado. Essa característica também faz com que o RMSE penalize modelos que possuem uma grande variação na magnitude dos erros calculados.

- Coeficiente de Determinação (R^2): Coeficiente que pode ser interpretado como a proporção das observações do AA que podem ser explicados pelos atributos utilizados (variando, geralmente, de 0 a 1). Importante ressaltar que o coeficiente não se relaciona com acurácia ou precisão, e sim com a correlação entre as observações reais do AA (y) e as suas respectivas previsões feitas pelo modelo (\hat{y}) (Kuhn & Johnson, 2013). Apesar de se conhecer diversas formas para efetuar seu cálculo (Kvålseth, 1985), o pacote “caret” adota uma abordagem mais direta, por meio do quadrado do valor obtido pelo cálculo da correlação entre y e \hat{y} (Kuhn, 2008).

Neste estudo, as três métricas apresentadas são calculadas e armazenadas para cada um dos modelos treinados, de forma a possibilitar a avaliação da performance dos mesmos. Entretanto, além da avaliação de performance entre modelos com o mesmo AA, resta ainda a necessidade de se comparar as perspectivas P1 (2 modelos finais, sendo 1 para FC e 1 para MN) e P2 (9 modelos finais, 5 para FC e 4 para MN), definidas na seção 3.4.4 para ser possível concluir a melhor abordagem dos dados para ML.

Desta forma, as métricas MAE e o RMSE não podem ser utilizadas para comparar modelos com AA diferentes, já que tais métricas são expressas conforme a unidade de medida e escala do AA analisado. Para resolver este problema, muito comum em estudos com dados que representam problemas reais a serem estudados com ML, como nos estudos de Bouman e van Laar (2006) e Feilhauer, Asner, Martin, e Schmidlein (2010), este estudo adotará suas versões normalizadas em

percentuais – $nMAE$ e $nRMSE$ – de forma a possibilitar a comparação que se deseja, bem como aprimorar interpretação dos resultados:

$$nMAE(\%) = \frac{MAE}{y_{max}-y_{min}} * 100 \qquad nRMSE(\%) = \frac{RMSE}{y_{max}-y_{min}} * 100$$

Para avaliar a performance do modelo utilizou-se o seguinte critério: primeiramente o modelo menor $nMAE$, e, em caso de empate, avaliação do $nRMSE$, seguido pelo R^2 . Finalmente, após a identificação do modelo de melhor performance, sua estrutura é armazenada em um arquivo para ser utilizado em futuras previsões e o algoritmo retorna para a etapa 2.4 da Figura 7, descrito na seção 3.4.7, para reiniciar o processo com o próximo AA, caso exista.

3.5 Conclusão

A seguir listam-se os aspectos mais relevantes da metodologia descrita:

- Foram apresentadas as características metodológicas do trabalho que auxiliam na definição da abordagem a ser descrita: um estudo de caso transversal, com filosofia positivista; abordagem indutiva e quantitativa; natureza exploratória-descritiva.
- A abordagem conceitual foi dividida em duas grandes fases: A primeira fase guarda estrita relação com o OE1 e OE2 e envolve etapas relacionadas a coleta, tratamento e armazenamento dos dados obtidos. A segunda fase se relaciona com os OE3 e OE4 e é composto por um algoritmo, que recebe os dados e aplica uma série de técnicas baseadas em ML para analisar os dados, treinar os modelos e avaliar os resultados.
- Foram apresentados os procedimentos de pré-processamento dos dados, as características dos modelos a serem treinados e por meio de quais métricas os mesmos são avaliados.

CAPÍTULO 4 APRESENTAÇÃO E DISCUSSÃO DOS RESULTADOS

4.1 Introdução

Este capítulo está estruturado da seguinte forma: a seção 4.2 apresentará os resultados obtidos por meio do treinamento manual de modelos, efetuado na etapa 2.8 da Figura 7 e detalhado na seção 3.4.12; a seção 4.3 apresentará os resultados obtidos por meio da plataforma de aprendizado automático, efetuado na etapa 2.9 da Figura 7 e detalhado na seção 3.4.13; a seção 4.4 se propõe a discutir os resultados de forma unificada; a seção 4.5 se baseará na aplicação de técnicas de IML para extrair dos modelos de melhor performance informações úteis para a tomada de decisão

e; por fim, a seção 4.6 apresentará um resumo das conclusões do capítulo. Cabe ressaltar também que os procedimentos adotados neste capítulo visam, sobretudo, atingir o OE3 “Validar um modelo, baseado em aprendizado de máquina, que possa explicar e prever substancialmente as variações nos gastos de FC e MN das instalações da MB, identificando suas potencialidades e limitações para futuros estudos”.

4.2 Resultados do Treinamento Manual de Modelos

Conforme descrito na seção 3.4.12, um total de 5 modelos foram treinados para cada execução do algoritmo. De acordo com a definição das perspectivas realizada na seção 3.4.4, um total de 11 diferentes AA serão analisados, o que resulta em um total de 55 modelos avaliados. Os resultados são apresentados nas Tabela 5 e Tabela 6, que relaciona as perspectivas com os modelos treinados e as métricas coletadas. Ao realizar uma breve avaliação direta da performance obtida, é possível traçar algumas conclusões iniciais, a seguir:

- Nenhum modelo demonstrou performance superior em todas as execuções, comprovando que as diferentes despesas e perspectivas possuem variabilidades distintas que são percebidas de forma diferente conforme as vantagens e desvantagens de cada modelo;
- SVR e RF foram os modelos mais bem sucedidos, entretanto XGB foi superior em P1 MN. RLM, apesar de ter sido incluído somente para fins de comparação, conseguiu ser superior a todos os modelos em P2 AGE e P2 ALG, evidenciando a preferência por uma abordagem linear para entender estes tipos de despesa;
- Considerando os resultados da Tabela 5 (despesas FC) e Tabela 6 (despesas MN), a perspectiva P1, que tem como propósito avaliar de forma conjunta os gastos de diferentes

Tabela 5 - Performance dos Modelos Manuais para Despesas FC em P1 e P2

	RLM			SVR			RF			GBM			XGB		
	nMAE (%)	nRMSE (%)	R ²	nMAE (%)	nRMSE (%)	R ²	nMAE (%)	nRMSE (%)	R ²	nMAE (%)	nRMSE (%)	R ²	nMAE (%)	nRMSE (%)	R ²
P1 FC	8,14	10,53	0,471	7,42	10,00	0,526	7,46	9,96	0,517	7,67	10,10	0,513	7,48	9,91	0,531
P2 GAS	10,54	14,27	0,364	9,94	14,13	0,380	10,93	14,68	0,336	10,44	14,35	0,340	10,47	14,30	0,351
P2 LIMP	11,13	14,01	0,300	10,95	13,81	0,314	10,78	13,71	0,304	11,19	13,80	0,315	10,88	13,61	0,337
P2 ENE	8,42	11,34	0,616	7,99	11,29	0,629	8,08	11,43	0,605	8,77	11,68	0,593	8,51	11,60	0,596
P2 AGEG	10,76	14,22	0,495	11,79	15,33	0,409	11,14	14,57	0,453	12,01	15,70	0,456	11,04	14,77	0,446
P2 COM	9,69	12,36	0,401	9,21	12,06	0,435	9,04	11,77	0,475	9,76	12,69	0,373	9,63	12,25	0,411

Tabela 6 - Performance dos Modelos Manuais para Despesas MN em P1 e P2

	RLM			SVR			RF			GBM			XGB		
	nMAE (%)	nRMSE (%)	R ²	nMAE (%)	nRMSE (%)	R ²	nMAE (%)	nRMSE (%)	R ²	nMAE (%)	nRMSE (%)	R ²	nMAE (%)	nRMSE (%)	R ²
P1 MN	8,87	11,56	0,269	8,19	10,96	0,346	8,38	11,08	0,339	8,57	11,21	0,314	8,15	10,74	0,371
P2 MNT	10,25	14,18	0,254	9,96	14,19	0,248	10,00	13,56	0,244	10,16	13,98	0,269	10,09	13,98	0,272
P2 XPTI	10,68	13,83	0,253	9,73	13,36	0,307	10,17	13,62	0,284	11,05	14,20	0,203	10,97	14,21	0,200
P2 ALG	12,52	16,35	0,122	13,28	17,61	0,021	13,58	17,74	0,058	12,95	16,96	0,057	12,85	16,77	0,090
P2 CPRP	14,62	18,45	0,311	13,48	17,93	0,342	13,11	17,36	0,380	14,00	18,15	0,322	14,10	17,94	0,332

categorias de despesa, obteve desempenho superior do que os modelos criados pela perspectiva P2, que tem como propósito avaliar as categorias de despesa isoladamente. Tal fato significa que o modelo foi capaz de aprender melhor sobre uma maior quantidade de dados de diferentes categorias de despesa do que sobre uma pequena quantidade de dados representando somente uma das categorias. Apesar do AA ter sido diferente em P1 e P2, as mensurações dos erros são comparáveis pelo fato da métrica estar normalizada e independente da unidade de medida e escala de magnitude de cada AA.

4.3 Resultados da Plataforma de Aprendizado Automático (H2O AutoML)

Conforme descrito na seção 3.4.13, a cada execução do algoritmo foi treinado um modelo adicional baseado no H2O AutoML. Considerando as perspectivas com diferentes AA, resultam-se em 11 modelos adicionais a serem analisados. Os resultados são apresentados nas Tabela 7 e Tabela 8, que relaciona as perspectivas com os modelos treinados e as métricas coletadas. Cabe ressaltar que, no caso dos *Stacked Ensembles*, é listado os principais modelos individuais que o compõem.

Tabela 7 - Performance dos Modelos H2O AutoML para as Despesas FC em P1 e P2

H2O AUTOML				
	nMAE (%)	nRMSE (%)	R ²	Modelo Vencedor
P1 FC	7,02	9,45	0,572	Stacked Ensemble (GBM;XGB;ANN)
P2 GAS	8,39	11,16	0,470	Stacked Ensemble (GBM;XGB;ANN)
P2 LIMP	10,95	13,72	0,307	Stacked Ensemble (MRL;RF;XGB)
P2 ENE	7,88	10,96	0,630	Stacked Ensemble (MRL;RF;XGB)
P2 AGE	10,61	13,57	0,520	Stacked Ensemble (MRL;RF;XGB)
P2 COM	9,25	12,15	0,41	Stacked Ensemble (GBM;XGB;ANN)

Tabela 8 - Performance dos Modelos H2O AutoML para as Despesas MN em P1 e P2

H2O AUTOML				
	nMAE (%)	nRMSE (%)	R ²	Modelo Vencedor
P1 MN	7,38	9,83	0,470	Stacked Ensemble (GBM;XGB;ANN)
P2 MNT	10,18	14,22	0,237	XGB
P2 XPTI	9,86	13,77	0,237	Stacked Ensemble (RLM;XGB;ANN)
P2 ALG	12,77	16,58	0,072	RLM
P2 CPRP	12,43	16,48	0,428	Stacked Ensemble (RLM;GBM;ANN)

Da mesma forma que na seção anterior, algumas breves conclusões podem tiradas dos resultados acima:

- Os *Stacked Ensembles* obtiveram performance superior em 9 dos 11 modelos treinados, confirmando o exposto na seção 3.4.14 sobre sua tendência de melhor desempenho. Entretanto, não ocorreu superioridade em termos de modelos individuais utilizados, onde pode ser observado que uma série de combinações distintas.
- Em um comportamento similar ao do treinamento manual, a perspectiva P1 obteve uma performance superior à de P2. Ou seja, a plataforma de aprendizado automático também foi capaz de desenvolver modelos que aprendem melhor sobre uma maior quantidade de dados de diferentes categorias de despesa do que sobre uma pequena quantidade de dados representando somente uma das categorias de cada vez.

4.4 Discussão dos Resultados Consolidados

Os resultados apresentados nas seções 4.2 e 4.3 representam abordagens diferentes para um mesmo problema e, portanto, devem ser consolidadas e discutidas em conjunto de forma a gerar informações úteis para o estudo. As Tabela 9 e Tabela 10 apresentam os melhores resultados de cada abordagem para cada perspectiva, identificando qual delas obteve melhor desempenho qual modelo foi utilizado.

Tabela 9 - Comparativo de Performance dos Modelos para Despesas FC em P1 e P2

	TREINAMENTO MANUAL			H2O AUTOML			Modelo Vencedor
	nMAE (%)	nRMSE (%)	R ²	nMAE (%)	nRMSE (%)	R ²	
P1 FC	7,42	10,00	0,526	7,02	9,45	0,572	Stacked Ensemble (GBM;XGB;ANN)
P2 GAS	9,94	14,13	0,380	8,39	11,16	0,470	Stacked Ensemble (GBM;XGB;ANN)
P2 LIMP	10,78	13,71	0,304	10,95	13,72	0,307	Random Forest (RF)
P2 ENE	7,99	11,29	0,629	7,88	10,96	0,630	Stacked Ensemble (MRL;RF;XGB)
P2 AGEG	10,76	14,22	0,495	10,61	13,57	0,520	Stacked Ensemble (MRL;RF;XGB)
P2 COM	9,04	11,77	0,475	9,25	12,15	0,41	Random Forest (RF)

Tabela 10 - Comparativo de Performance dos Modelos para Despesas MN em P1 e P2

	TREINAMENTO MANUAL			H2O AUTOML			Modelo Vencedor
	nMAE (%)	nRMSE (%)	R ²	nMAE (%)	nRMSE (%)	R ²	
P1 MN	8,15	10,74	0,371	7,38	9,83	0,470	Stacked Ensemble (GBM;XGB;ANN)
P2 MNT	9,96	14,19	0,248	10,18	14,22	0,237	Support Vector Regression (SVR)
P2 XPTI	9,73	13,36	0,307	9,86	13,77	0,237	Support Vector Regression (SVR)
P2 ALG	12,52	16,35	0,122	12,77	16,58	0,072	Regressão Linear Múltipla (RLM)
P2 CPRP	13,11	17,36	0,380	12,43	16,48	0,428	Stacked Ensemble (MRL;RF;XGB)

Ao analisar os resultados de forma consolidada, foi possível traçar as seguintes conclusões:

- Dos 11 AA analisados, 6 obtiveram desempenho superior pela abordagem H2O AutoML, enquanto 5 obtiveram desempenho superior em um dos modelos treinados manualmente por meio do pacote “caret”.
- Entretanto, como a perspectiva P1 obteve os menores índices de erro em comparação com a perspectiva P2, ao analisar somente seus dois modelos (P1-FC e P1-MN) identifica-se a superioridade da abordagem H2O AutoML, que produziu *Stacked Ensembles* capazes de gerar a melhor performance obtida neste estudo. Interessante notar que os melhores resultados utilizaram a mesma combinação de modelos individuais no processo de aprendizado: GBM, XGB e ANN.
- Um dos motivos que podem explicar tal comportamento é a utilização de ANN. Este método não foi aplicado pela abordagem manual devido à sua complexidade, entretanto trata-se de uma poderosa técnica de regressão capaz de identificar padrões não lineares, mas que requer uma grande quantidade de dados para ter resultados satisfatórios (Kuhn & Johnson, 2013; Markham & Rakes, 1998).
- O nível de erro apresentado pelo nMAE (%), nRMSE (%), e R^2 , apesar de não serem demasiadamente grandes, podem fazer com que os modelos não sejam suficientemente precisos, em um primeiro momento, para a efetiva adoção do modelo como um apoio à tomada de decisão. Entretanto, cabe ressaltar que, conforme exposto na seção 3.4, o foco deste estudo não é de encontrar a melhor performance possível, e sim validar a utilização de ML para solução do problema exposto. Além disso, considerando que a metodologia descrita no Capítulo CAPÍTULO 3 foi desenvolvida com o propósito de possibilitar estudos futuros com novos atributos e outros problemas de interesse da organização e também que alguns atributos coletados neste trabalho não foram utilizados por problemas de compatibilidade, conclui-se que as Fases 1 (Coleta de Dados) e 2 (Aprendizado de Máquina) guardam o potencial de significativa melhora dos resultados obtidos.
- Outra explicação complementar para os níveis de erro apresentados é a existência de incertezas não facilmente representáveis em atributos no ambiente onde a organização se insere. Questões políticas e econômicas, cortes e contingenciamentos orçamentários, perfil dos gestores em todos os níveis da organização, novos regulamentos, perfil dos usuários utilizadores de recursos geradores dos gastos estudados, entre outros.

Por fim, para ilustrar a performance obtida no modelo final da perspectiva P1-FC, a Figura 19 apresenta um gráfico onde a linha vermelha representa as observações reais (y) e os pontos azuis

representam as previsões feitas pelo modelo treinado (\hat{y}). Para facilitar o entendimento, y foi ordenado de forma crescente de forma a evidenciar algumas características interessantes do resultado:

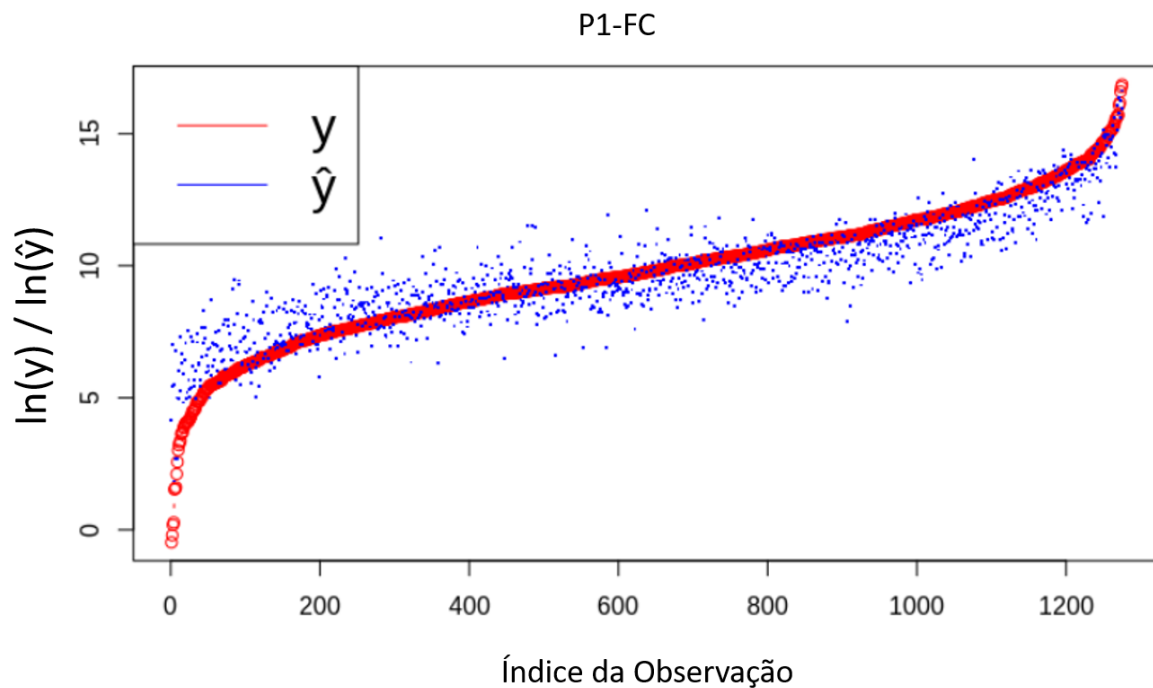


Figura 19 - Observações Reais x Valores Previstos para P1-FC

- É possível notar que, de uma forma razoável, \hat{y} acompanha a tendência de crescimento de y . Entretanto, quando o valor de y é pequeno, existe a tendência de \hat{y} ser maior e quando y assume valores maiores, existe tendência de \hat{y} ser menor.
- A dispersão dos pontos azuis em torno da linha vermelha é a representação visual do erro que o modelo possui. A inclusão de novos atributos e o refinamento das técnicas aplicadas tem o potencial de reduzir tal medida.

4.5 Resultados adicionais por meio de *Interpretable Machine Learning (IML)*

Conforme já exposto na seção 3.3, uma das maiores limitações de uma abordagem baseada em ML é a falta de transparência sobre o funcionamento interno de um modelo treinado, que pode ser resumido pela falta de interpretabilidade. Muitas das vezes, para o decisor, não basta confiar nas previsões a serem realizadas pelo modelo, é necessário entender os motivos pelos quais a previsão foi feita, não apenas por curiosidade, mas para que desta forma conhecimento sobre seu negócio seja efetivamente gerado (Du et al., 2019).

Esta seção se propõe a aplicar técnicas de IML sobre os melhores resultados encontrados para extrair informações úteis para a tomada de decisão. Desta forma, somente os dois modelos finais da

perspectiva P1 serão analisados. As técnicas de IML são divididas em duas categorias: aquelas que são dependentes de algum modelo específico e as funcionam independente do modelo a ser estudado. Este estudo se focará na aplicação de técnicas que independem do modelo por duas razões:

- O acréscimo de novos dados, ou estudo de outros problemas de interesse da organização pode vir a dar resultados medidos por modelos diferentes, o que comprometeria a técnica de IML adotada; e
- Os melhores resultados deste estudo foram obtidos por meio de *Stacked Ensembles* que, por definição, consistem em um agregado de modelos individuais. Entretanto, tal característica dificulta sua interpretação, uma vez que somente são suscetíveis a técnicas independentes de modelo.

4.5.1 Importância dos Atributos por Permutação (*Permutation Feature Importance*)

Técnica idealizada por Fisher, Rudin, & Dominici (2019), com base no conceito de Importância de Atributos introduzida por Breiman (2001) durante o desenvolvimento do modelo *Random Forest* (RF). Seu funcionamento consiste em avaliar a importância de um determinado atributo para a performance final do modelo por meio do embaralhamento de suas observações, na tentativa de afetar seu relacionamento com o AA. Se, após o embaralhamento, o erro produzido pelo modelo aumenta, o atributo analisado é considerado importante (Boehmke & Greenwell, 2019).

Neste estudo, os modelos finais da perspectiva P1 foram avaliados para identificação dos atributos mais relevantes, cuja mensuração da importância é efetuada em uma escala que representa a força da sua contribuição para a previsão do AA. A técnica foi implementada por meio do pacote “vip” (*Variable Importance Plots*) da linguagem R. A Figura 20 apresenta um gráfico com os 10 atributos mais importantes nos modelos da perspectiva P1. Algumas conclusões podem ser efetuadas por meio da análise dos resultados:

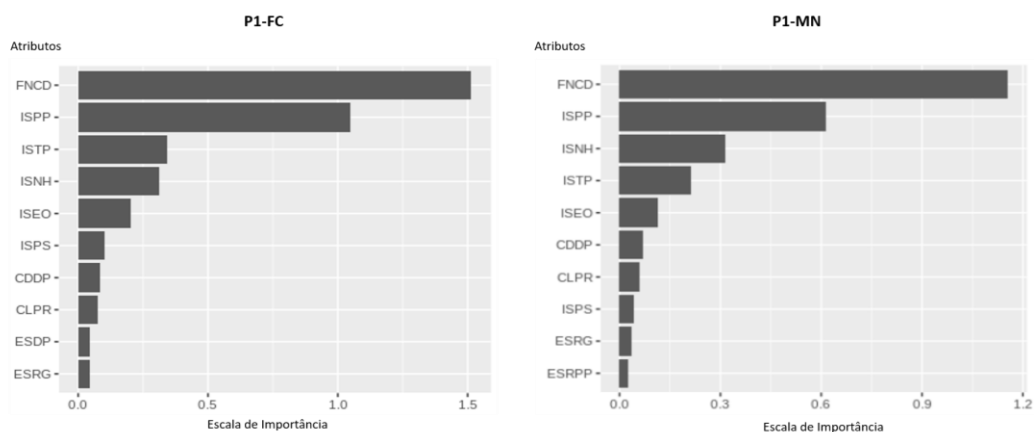


Figura 20 - Importância dos Atributos nos Modelos da Perspectiva P1

- De uma forma geral, os 5 primeiros atributos mais importantes de cada modelo são iguais (FNCD, ISPP, ISTP, ISNH, ISEO), possuindo apenas ligeiras diferenças na ordem. Estes atributos representam, além da categoria da despesa, informações relacionadas a cada instalação, de forma específica (quantidade de pessoas, tipo de instalação, nível hierárquico, localização na estrutura organizacional).
- Nos 5 atributos restantes a escala de importância é significativamente reduzida, e apenas 3 atributos se repetem nos dois modelos (ISPS, CDDP e CLPR). Desta forma, além da característica da instalação de ser prestadora de serviço, atributos de caráter social e geográfico aparentam ficar em segundo plano em termos de relevância para a previsão do AA.
- O fato de FNCD ser categórica e também o atributo mais importante encontrado indica que algumas categorias de despesa são cruciais para a formação da previsão do AA. Entretanto a técnica não é capaz de detalhar quais categorias são relevantes e quais não são, o que compromete a interpretabilidade deste resultado.
- Já no caso de ISPP, que é numérica e o segundo atributo mais importante, os resultados mostram que a quantidade de pessoal alocado em cada instalação é um fator substancialmente relevante na alocação dos gastos FC/MN. Desta forma, os decisores devem estar atentos que qualquer ação que tem como característica a movimentação de pessoal entre instalações tem o potencial de alterar a composição destes gastos. Apesar de ter sido identificado que os atributos sociais e geográficos não sejam de importância relevante, a escolha cuidadosa das instalações de origem e destino de uma eventual movimentação permanece essencial, pois atributos como tipo de instalação, hierarquia e posição na estrutura da organização podem explicar uma fração significativa dos associados.

As conclusões suportam que o modelo pode ser utilizado no dimensionamento dos gastos oriundos de decisões que envolvam, por exemplo, a movimentação urgente de tropas pelo território nacional, ou a ativação/desativação de instalações da organização. Nestes casos, os modelos são capazes de gerar estimativas de redução ou acréscimo nos gastos com base na saída ou chegada de pessoal na instalação, levando em consideração as características da mesma.

Por fim, os apontamentos acima são capazes de produzir previsões que podem ser utilizadas como um critério adicional a ser levado em consideração ao longo do processo de tomada de decisão da organização. Além disso, o conhecimento dos atributos mais relevantes pode orientar futuros

esforços de coleta de dados em direção a uma melhoria mais significativa na performance dos modelos. Por exemplo, os atributos FNMQ (metros quadrados de área construída) e FNPT (patrimônio), removidos por problemas de compatibilidade, possuem características semelhantes a ISPP: são numéricos, individualizados para cada instalação e representantes do seu tamanho e/ou complexidade de cada uma. Portanto, este comportamento sugere que os atributos removidos deste estudo poderiam aprimorar substancialmente a performance dos modelos treinados.

4.5.2 *Partial Dependence Plot (PDP)*

PDP é uma técnica que auxilia no entendimento do efeito que um determinado atributo possui para a performance preditiva do modelo. Idealizado por Friedman (2001). De uma forma resumida, sua aplicação permite entender como a previsão do AA varia quando os valores do atributo são modificados, sem deixar de considerar o efeito de todos os outros atributos para o resultado final. A previsão é efetuada diversas vezes com vários valores modificados e o resultado é apresentado em um gráfico que representa a média dos resultados obtidos. Além de serem fáceis de implementar, são bastante intuitivas e eficientes no propósito de acrescentar interpretabilidade aos resultados.

A técnica foi implementada por meio do pacote “pdp” da linguagem R, tendo como foco a análise do relacionamento dos 5 atributos mais relevantes (FNCD, ISPP, ISTP, ISNH e ISEO) com as previsões efetuadas pelos modelos (\hat{y}), para cada um dos modelos da perspectiva P1. Cabe ressaltar que os resultados de atributos categóricos foram ordenados de forma crescente para uma melhor visualização e que \hat{y} representa o atributo de gasto (FNCT) transformado pela função logarítmica. Algumas conclusões podem ser efetuadas por meio da análise dos resultados apresentados nas Figuras 21 e 22.

- ISPP: A análise dos dois modelos demonstra resultados similares, evidenciando uma relação aproximadamente linear da quantidade de pessoas alocadas em uma instalação e os seus gastos em FC/MN. O relacionamento linear fica mais evidente em instalações com mais de 20 pessoas aproximadamente, o que sugere que, em instalações muito pequenas, a variabilidade é baixa em virtude dos custos fixos associados.

P1-FC

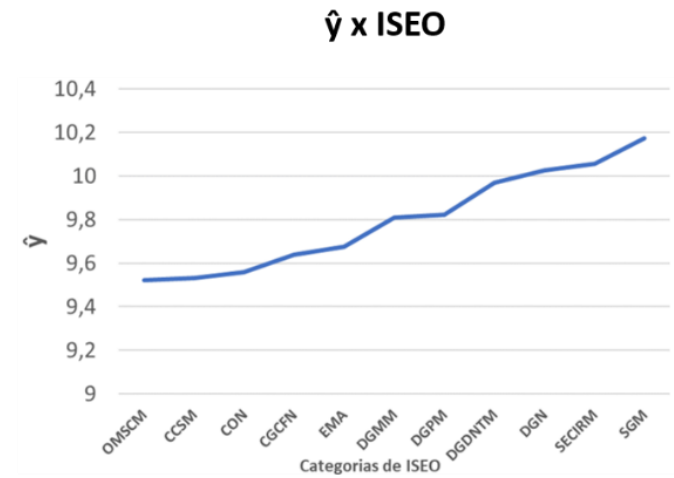
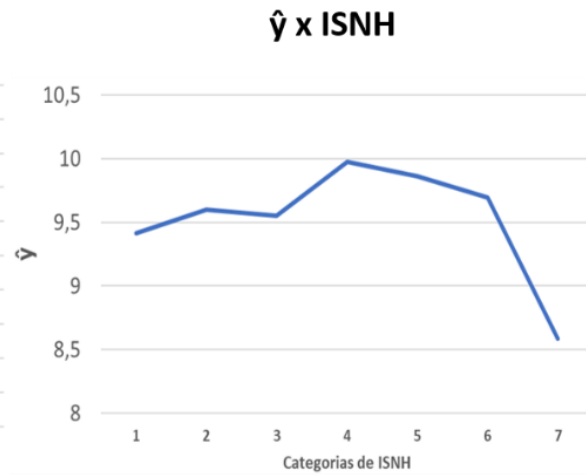
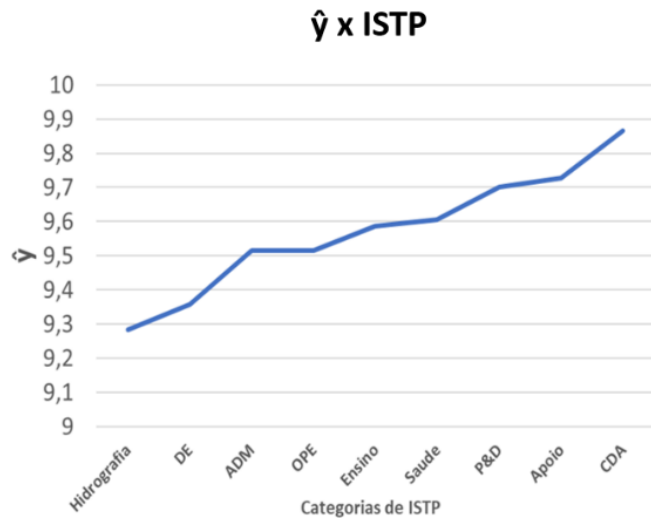
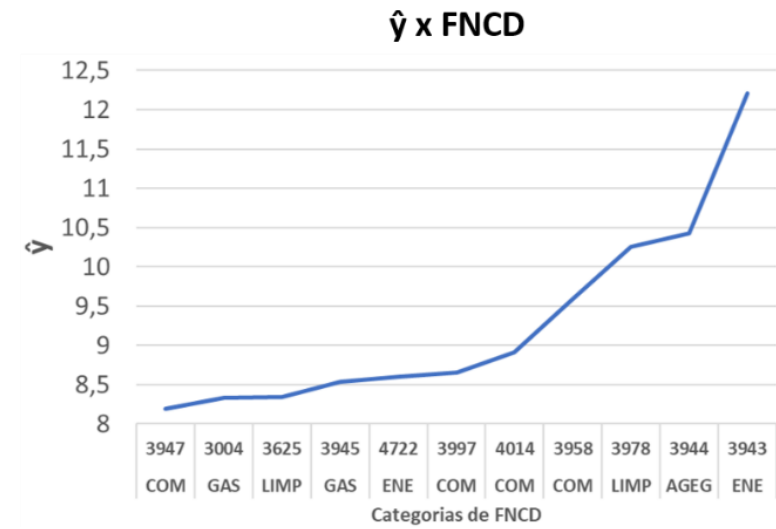
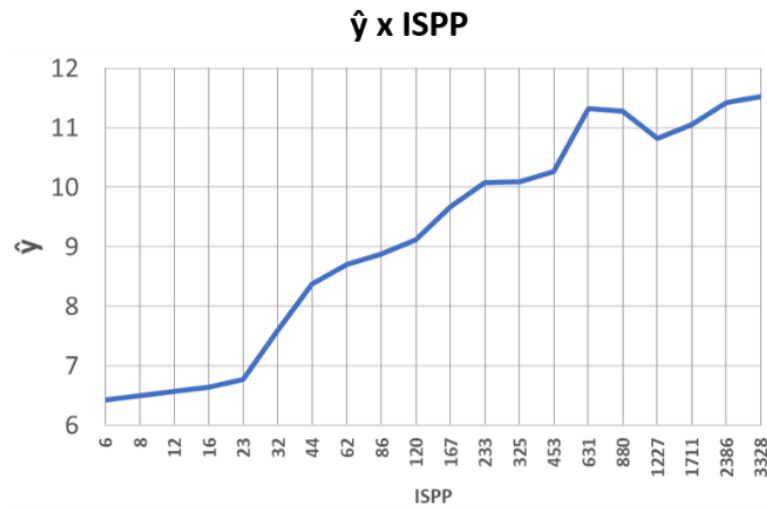
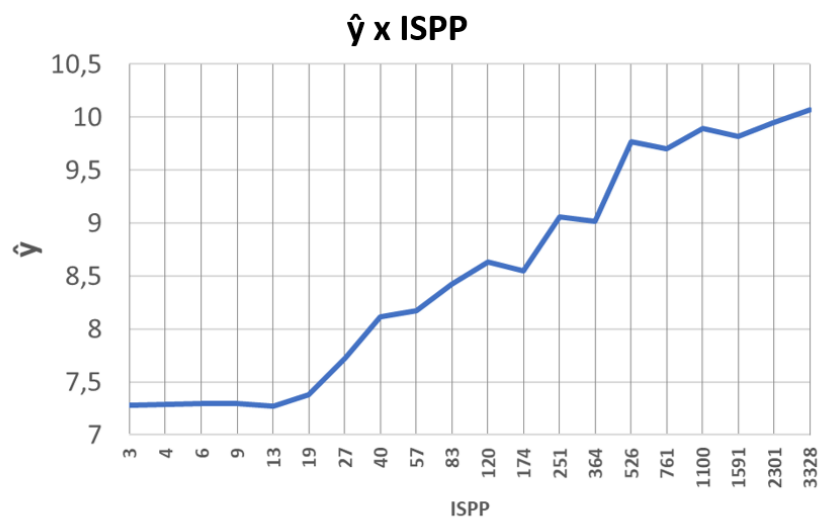
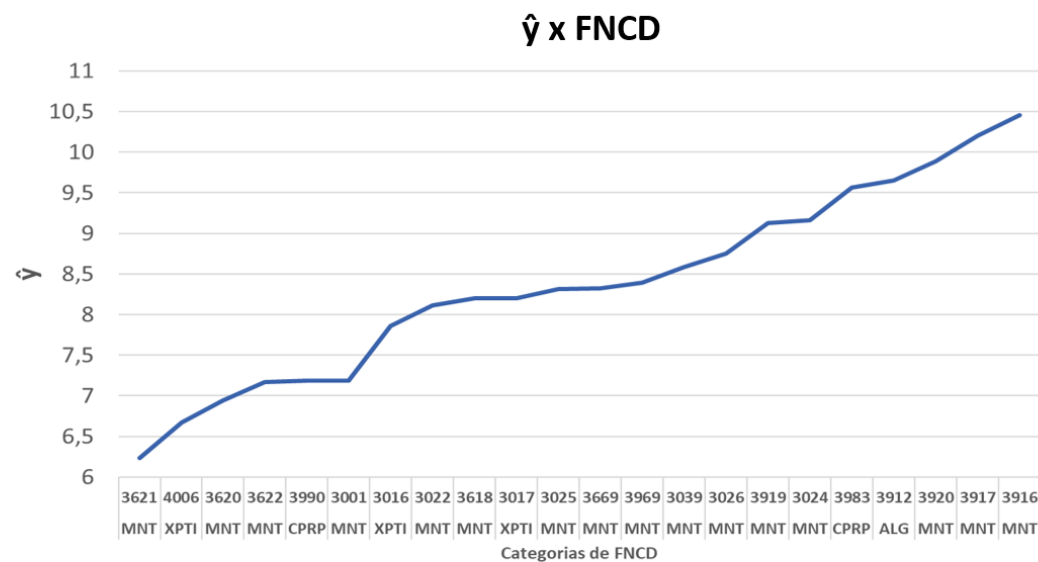


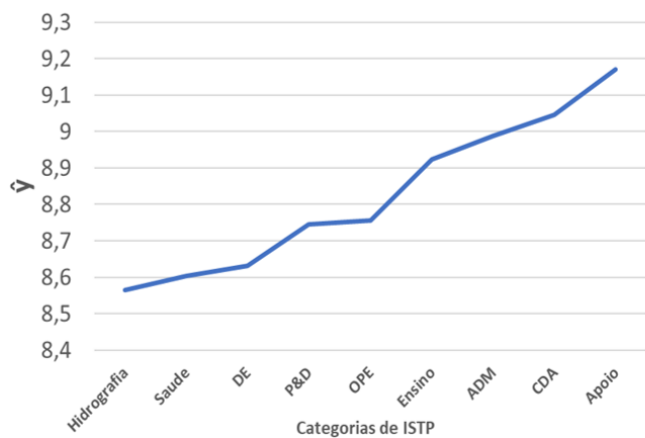
Figura 21 - Resultados da técnica PDP nos atributos relevantes de P1-FC



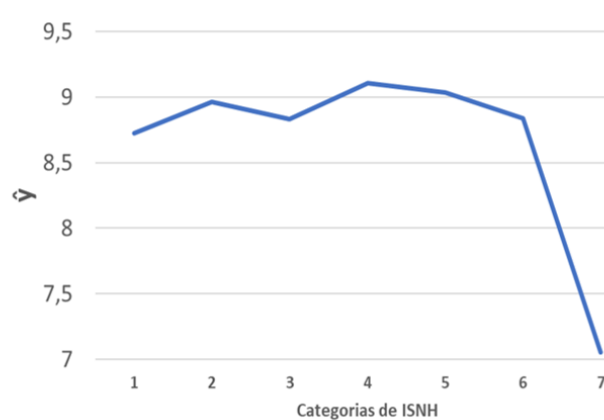
P1-MN



ŷ x ISTP



ŷ x ISNH



ŷ x ISEO



Figura 22 - Resultados da técnica PDP nos atributos relevantes de P1-MN

FNCD: Por ser um atributo categórico, o resultado do PDP deve ser entendido como uma análise de suas classes, em termos de sua relevância para a composição da previsão. No modelo P1-FC, as primeiras 3 classes (conforme listagem presente no Apêndice II) obtiveram resultados significativamente superiores aos demais: 3943 “Serviços de Energia Elétrica”, 3944 “Serviços de Água e Esgoto” e 3978 “Limpeza e Conservação (PJ)”. O fato dos gastos de água/esgoto e energia elétrica terem sua importância comprovada suporta a sugestão de que os resultados poderiam ser substancialmente melhores caso a organização obtivesse dados que representassem o consumo de tais recursos, medidos em KWh, no caso da energia elétrica, e m³, no caso de água/esgoto. Tais valores eliminariam distorções provocadas pela variação do preço a ser pago por tais serviços ao longo do território brasileiro.

Já Em P1-MN, as classes observam uma crescente de importância de comportamento linear, onde as 3 mais importantes são 3916 “Manutenção e Conservação de bens imóveis (PJ)”, 3917 “Manutenção e Conservação de máquinas e equipamentos” e, 3920 “Manutenção e Conservação de bens móveis de outras Naturezas”.

- ISNH: O resultado do PDP deve ser entendido por meio do mesmo raciocínio seguido para o atributo FNCD. Ambos os modelos apresentaram resultados similares, onde a única variabilidade significativa é para o último nível da hierarquia, que possui valores previstos menores. Uma breve consulta aos dados de origem retorna que as instalações designadas neste último nível são os navios da Marinha do Brasil, conforme identificados no Apêndice III, possuindo a maior quantidade de instalações em nível superior de toda a Estrutura Organizacional.

Os navios, enquanto atracados, são apoiados por uma base ou estaleiro, que fornecem água, energia, materiais e quaisquer outros recursos que sejam necessários; já no mar, a energia é fornecida por geradores internos, a água é consumida dos tanques abastecidos em terra, mas também complementada por técnicas de dessalinização da água, como a osmose reversa, presente em diversos navios da organização. Tal peculiaridade faz com que os resultados não possam ser interpretados de forma a sugerir que os navios possuem gastos substancialmente menores e sim, que seu comportamento se difere das demais instalações analisadas, sendo difícil a integral associação de seus gastos, que podem estar fragmentados nas bases que adquirem os recursos necessários, como combustíveis, água, sobressalentes para manutenção, entre outros.

Tais apontamentos sugerem que essa peculiaridade pode ter sido crucial para que este atributo tenha sido considerado como relevante pelo modelo, já que é o único atributo capaz de isolar os navios das demais instalações, sendo, portanto, a única forma que o modelo possui de diferenciá-los.

- ISEO e ISTP: Os resultados da técnica PDP para estes dois atributos foram bastante similares.

Em P1-FC e P1-MN, apesar das classes destes atributos terem evidenciado rankings ligeiramente diferentes, nota-se que a escala de \hat{y} é bastante reduzida, variando menos de 1 unidade, o que sugere que tais atributos são utilizados pelo modelo como uma espécie de “ajuste fino”, provendo pequenos ajustes com base em suas classes para melhorar a performance final dos modelos.

A técnica PDP também pode ser utilizada para entender a contribuição conjunta de um par de atributos sobre o AA. Dessa forma, é possível aprimorar o nível das interpretações a serem extraídas do modelo para apoio a tomada de decisão. A Figura 23 apresenta um PDP que envolve a contribuição de ISPP e FNCD para as previsões do AA em P1-FC. Esta forma de visualização facilita a compreensão de como uma variação na quantidade de pessoas em uma determinada instalação (ISPP) afeta as despesas de FC para cada categoria de gasto. Além das curvas, são exibidas as linhas de tendência com sua respectiva equação. Algumas conclusões adicionais podem ser extraídas, a seguir:

- A taxa de crescimento de todas as categorias de gasto em FNCD parece aumentar quando $\log(\text{ISPP}) > 3,5$ (aprox. 30 pessoas) e diminuir quando $\log(\text{ISPP}) > 7$ (aprox. 1090 pessoas).
- ENE é a categoria mais relevante para qualquer valor do FCPP.
- Ao analisar a inclinação das linhas de tendência (os valores em negrito na Figura 23), entende-se que COM é um pouco menos sensível ao aumento de pessoal do que o restante das categorias.

4.6 Conclusão

A seguir listam-se os aspectos mais relevantes apresentados neste capítulo:

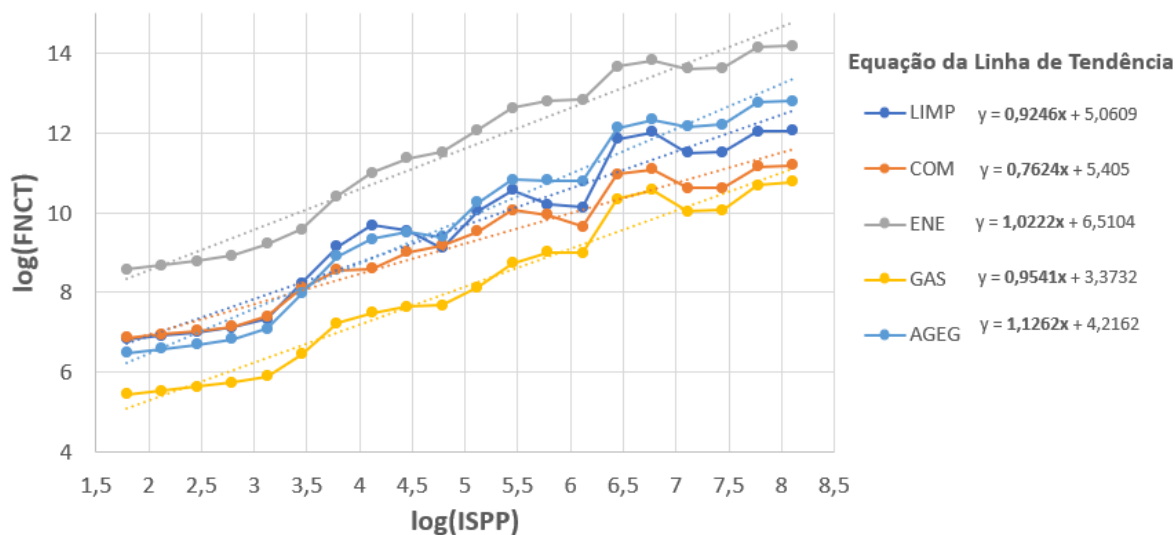


Figura 23 - PDP de ISPP e FNCD sobre o as previsões do AA (FNCT)

- De uma forma geral, os modelos criados a partir da perspectiva P1, que considera as despesas agregadas, obtiveram performance superior aos modelos da perspectiva P2, que considera cada tipo de despesa de forma isolada. A comparação entre diferentes AA foi possível por meio da utilização das métricas de performance normalizadas definidas na seção 3.4.14;
- Dentre os resultados obtidos pelo treinamento manual de modelos, os modelos baseados em SVR e RF foram os mais bem sucedidos em 8 dos 11 AA definidos na seção 3.4.4. RLM obteve melhor performance em 2 casos e XGB em apenas 1 caso;
- Entretanto, os resultados obtidos pela plataforma de aprendizado automática foram superiores em 6 dos 11 AA definidos. Entre os 6 casos estão P1 MN e P1 FC, que foram então considerados os melhores resultados obtidos por este estudo, e foram submetidos a técnicas de IML para identificação dos atributos mais relevantes e descrição de como estes se relacionam com o AA. A Tabela 11 resume as principais informações dos modelos finais.

Tabela 11 - Informações sobre os modelos de melhor performance

AA	nMAE (%)	nRMSE (%)	R ²	Modelo Criado	Atributos Relevantes
P1 MN	7,38	9,83	0,470	Stacked Ensemble	FNCD, ISPP, ISTP, ISNH, ISEO
P1 FC	7,02	9,45	0,572	Stacked Ensemble	FNCD, ISPP, ISNH, ISTP, ISEO

CAPÍTULO 5 SUGESTÃO DE IMPLEMENTAÇÃO

5.1 Introdução

Este capítulo tem como propósito estabelecer diretrizes para a implementação, de forma sistêmica, da abordagem conceitual desenvolvida neste estudo. Tal necessidade reside no fato que uma parcela significativa do esforço deste trabalho foi direcionada na validação do ML como uma abordagem eficaz no entendimento dos gastos FC/MN da organização e, portanto, não objetivou desenvolver modelos com a melhor performance possível, o que certamente pode ser alcançado por meio do refinamento das técnicas aplicada e pela utilização de uma quantidade mais diversa de atributos.

Desta forma, a sugestão de implementação envolve alguns aspectos relacionados com a Tecnologia da Informação e Comunicação (TIC) e tem como propósito permitir que a organização se beneficie com a abordagem desenvolvida, possibilitando a continuação deste e de outros estudos, sem que haja a necessidade de profundo conhecimento prévio sobre os mesmos e com uma menor participação de especialistas na área de tecnologia/gestão, uma vez que muitos dos passos necessários são automatizados. A sua adoção pela organização permite, inclusive, a reprodução deste estudo com menores restrições de tempo, dados e processamento. Conforme citado na seção 3.4, algumas premissas foram consideradas ao longo deste trabalho para que tal sugestão se tornasse possível: além da utilização de software livre, para que sua implementação não implique em custos adicionais de licenciamento, o próprio desenvolvimento da metodologia priorizou uma abordagem focada na reutilização e flexibilidade, de forma que seja facilmente aplicável em outros contextos.

Finalmente, cabe ressaltar também que as sugestões efetuadas neste capítulo visam, sobretudo, atingir o OE4 “Avaliar se o modelo pode ser implantado de forma sistêmica dentro da organização, com o apoio de ferramentas tecnológicas de código aberto (software livre), de preferência”.

5.2 Protótipo da *Framework* de Aprendizado

Conforme apresentado na seção 2.2.3, as organizações buscam aprimorar sua capacidade analítica, por meio do aumento do uso extensivo de dados, análises estatísticas e quantitativas, modelos preditivos e explicativos, além de uma gestão baseada em fatos para guiar decisões e ações. Normalmente, para que softwares com capacidades analíticas possam ser utilizados de forma eficaz, é

necessário que sejam apoiados por um processo capaz de oferecer dados na melhor qualidade possível.

A Figura 23 apresenta uma ilustração básica de um processo ETL (*Extraction, Transform, Load*), a partir da qual alguns aprimoramentos serão apontados e discutidos para que a estrutura fundamental que possibilite o aprendizado de máquina seja, enfim, obtida.

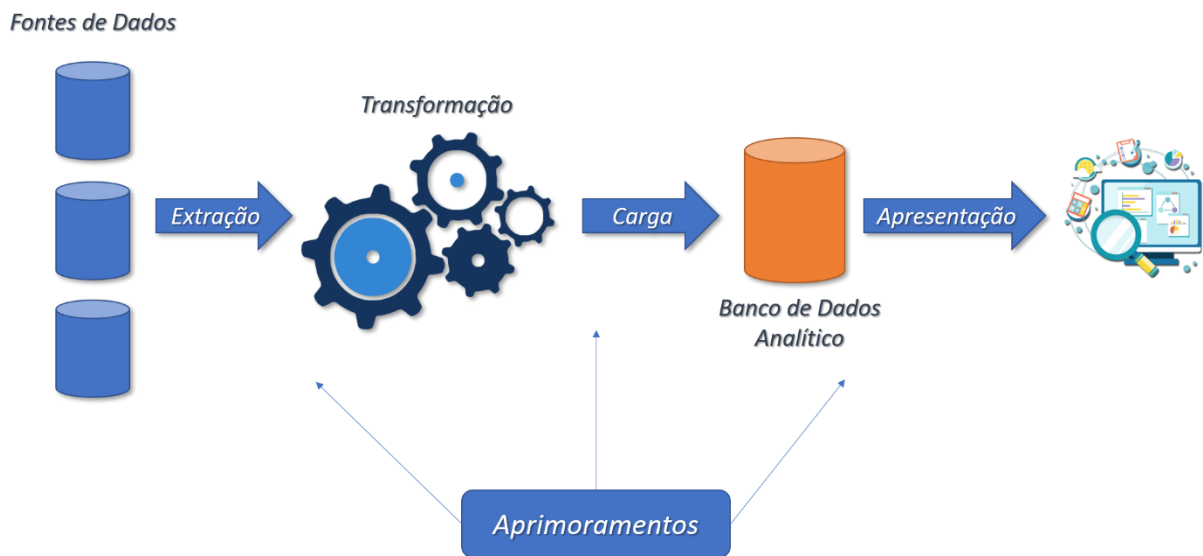


Figura 24 - Ilustração de Processo ETL

5.3 Fontes de Dados e Extração

O processo se inicia com o acesso a fontes de dados externas, o que possibilita a criação de um procedimento que possa extraí-los para sua efetiva utilização. Com relação à contribuição das fontes de dados para o processo de aprendizado, é necessário ter em mente que o propósito é a característica do objeto em estudo que possa aprimorar a performance do modelo, por meio de melhorias no entendimento da variabilidade dos dados sendo analisados. Ou seja, mesmo que o dado não tenha utilidade direta na etapa de apresentação, o mesmo ainda pode ser utilizado para obter previsões mais precisas. Um exemplo direto encontrado neste estudo é o caso dos atributos relacionados a patrimônio (FNPT) e metros quadrados (FNMQ) de cada instalação, que não puderam ser utilizados por problemas de contabilidade. Neste ponto, o interesse e o comprometimento da organização vão guiar suas ações em torno de efetuar mudanças para uma orientação a dados mais efetiva.

Para o processo de extração, com o objetivo de estabelecer uma consistência e regularidade, sugere-se que a organização estabeleça relações ou protocolos com gestores de fontes de dados externas, de forma tal que o fornecimento possa ser realizado por meio de um processo automatizado.

A coleta manual, como a que foi efetuada neste estudo, deve ser utilizada apenas para fins de estudos que possuam o propósito de analisar, de forma prévia, a utilidade do dado a ser obtido.

5.4 Transformação e Carga

As ações relacionadas à transformação dos dados devem ser repensadas de forma que passe a apoiar também dados destinados ao processo de aprendizado. Por exemplo, quando o propósito é a simples apresentação dos dados, valores ausentes não possuem valor e, portanto, podem ser inseridos procedimentos de remoção do registro quando identificado um destes casos. No entanto, ao se considerar o processo de aprendizado, um procedimento como este pode até piorar a performance do modelo final, uma vez que valores ausentes podem ser tratados, conforme descrito na seção 3.4.6, para que o restante dos valores associados ao registro identificado possa participar do processo de aprendizado. Distribuições também podem ser analisadas e transformadas de forma prévia, conforme a necessidade de cada atributo coletado.

No que se refere à carga dos dados, há de se ressaltar que a arquitetura de dados também deverá ser reformulada para comportar os resultados da transformação. Primeiramente, de forma a tornar evidente a compatibilidade dos novos dados que estão sendo inseridos com os antigos já utilizados; tal compatibilidade pode ser entendida como a capacidade que um novo atributo possui de se relacionar com todos os outros por meio de um campo chave. Finalmente, deverão ser designados os espaços físicos (em caso de necessidade) e virtuais (em termos de memória e tabelas de alocação) necessários a inserção de novos dados.

5.5 Banco de Dados Analítico e Apresentação

Uma vez armazenados todos os dados a serem utilizados no processo de aprendizado, é necessário que as etapas listadas na Fase 2 da abordagem conceitual descrita na seção 3.4 sejam executadas. A sugestão é de que o algoritmo apresentado seja executado de forma periódica (a cada 24 horas, por exemplo), ou sempre que novos dados sejam inseridos, em um horário com baixa demanda, com base em algumas preferências previamente definidas pela organização, tais como:

- Atributos-Alvo (AA) a serem analisados;
- Tempo limite de treinamento da plataforma automática;
- Preferências relacionadas ao particionamento dos dados (seção 3.4.8); e
- Preferências de métricas e gráficos a serem produzidos.

Com as informações acima, o algoritmo pode ser executado e seus resultados armazenados para serem fornecidos quando solicitados pela etapa de apresentação, que também deverá ser adaptada para possibilitar que o usuário possa usufruir dos resultados da aprendizagem. A Figura 24 ilustra o funcionamento desta etapa.

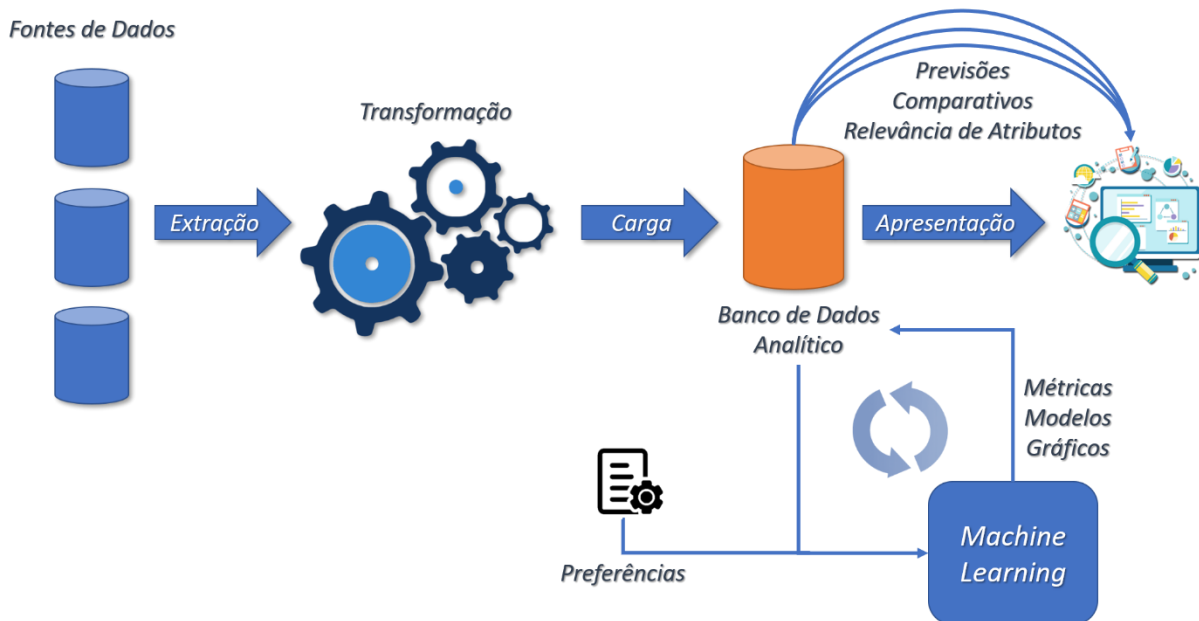


Figura 25 - Implementação da rotina para Aprendizado de Máquina

5.6 Potenciais Benefícios

Conforme descrito na seção 2.2.3, baseado nos estudos de Davenport e Harris (2017) a respeito da capacidade analítica das organizações, as sugestões desse capítulo objetivam a implementação sistemática da abordagem conceitual presente neste estudo para dotar a Marinha do Brasil de capacidade analítica preditiva, prescritiva e autônoma que lhe possibilite coletar conhecimento que transcenda ao que foi gerado neste estudo, ao abordar diferentes situações e contextos que possam ser de interesse. Entretanto, de acordo com os autores mencionados, enquanto tais capacidades estão diretamente relacionadas com a obtenção de vantagens competitivas no contexto de mercado, quando se trata de uma organização integrante da Administração Pública o foco é o aprimoramento da eficiência na alocação de recursos para o atendimento das demandas da sociedade. A Marinha do Brasil, portanto, tende a contribuir em tal aspecto ao utilizar ferramentas capazes de apoiar a decisão na gestão de seus recursos, visando a identificação de oportunidades onde seja possível gastar menos sem comprometer sua prontidão e capacidade operativa.

Outro potencial benefício da implementação da abordagem conceitual deste estudo é usufruir, mesmo que superficialmente, das vantagens de orientar a decisão com base em dados. Neste estágio

inicial, a geração de percepções que sejam suficientemente úteis pode auxiliar a organização a superar as dificuldades expostas na seção 2.2.2, relacionadas ao tratamento do dado como um ativo organizacional essencial para formulação de sua estratégia. Quanto maior for a importância do dado dentro da organização, mais atributos podem ser mensurados e tratados com maior qualidade, resultando em melhor performance final dos modelos treinados.

CAPÍTULO 6 CONCLUSÕES

Este capítulo apresentará as conclusões deste estudo por meio de uma breve contextualização dos objetivos definidos ao longo do mesmo. Após uma breve introdução que contextualiza e situa o problema de pesquisa, que tem origem na necessidade de se considerar as especificidades e diversidades das instalações da MB no entendimento das despesas de FC e MN da organização, foi definido como objetivo geral do estudo “desenvolver um modelo para apoio à tomada de decisão na gestão dos gastos da Marinha do Brasil”. Sua relevância foi confirmada por meio da revisão da literatura exposta na seção 2.3, por meio da qual conclui-se que são poucas as pesquisas que se propõem a apoiar a decisão na gestão do orçamento na área de Defesa. Seu atingimento está diretamente relacionado com a definição e desenvolvimento dos objetivos específicos, resumidos a seguir.

O primeiro objetivo específico “OE1 - Identificar, coletar e analisar os dados atualmente disponibilizados que representem atributos das instalações da organização e que possam auxiliar no desenvolvimento do modelo” foi alcançado, primeiramente, pela revisão da literatura relacionada a importância e relevância nos dados nas organizações, exposto nas seções 2.2.1 e 2.2.4, que auxiliou no processo de identificação dos dados a serem coletados. Posteriormente, as seções 3.4.1 e 3.4.2 descrevem como os dados identificados foram coletados de suas fontes de origem e analisados para sua utilização no desenvolvimento do modelo.

O atingimento do segundo objetivo específico “OE2 - Desenvolver uma infraestrutura capaz de gerenciar os dados coletados de forma eficaz, possibilitando o relacionamento entre os mesmos, bem como diversos tratamentos e abordagens” iniciou-se também na revisão da literatura, uma vez que as seções 2.2.2 e 2.2.3 exploram estudos que descrevem o tratamento a ser oferecido aos dados, em termos de gestão, integração e cultura organizacional para que possam resultar em benefícios e significantes transformações. Assim, as seções 3.4.3 e 3.4.4 apresentam a arquitetura de dados desenvolvida e como ela foi aplicada para gerar diferentes perspectivas de dados, utilizadas como entrada do processo de aprendizado.

Já o terceiro objetivo específico “OE3 - Validar um modelo, baseado em aprendizado de máquina, que possa explicar e prever substancialmente as variações nos gastos de FC e MN das instalações da MB, identificando suas potencialidades e limitações para futuros estudos” consiste na maior parte da pesquisa efetuada. Primeiramente, a revisão da literatura descrita na seção 2.2.5 auxiliou no entendimento sobre o que é o aprendizado de máquina, quais as razões de sua utilização,

sua relevância ao longo do tempo e também apresentam pesquisas com aplicações nas mais diversas áreas de estudo. Posteriormente, com foco em na metodologia adotada, a seção 3.3 descreve seu funcionamento e aponta como os modelos gerados por aprendizado de máquina podem ser utilizados na solução do problema de pesquisa, oferecendo potencialidades e limitações de sua aplicação. Em seguida, a seção 3.4 descreve todas as etapas do processo de aprendizado, dividido em duas fases que cobrem desde a extração dos dados das fontes de origem até o cálculo das métricas utilizadas para avaliação da performance dos modelos treinados. Por fim, o Capítulo 4 apresenta e discute os resultados obtidos, além de aplicar técnicas adicionais para interpretar algumas características dos modelos de melhor performance.

Com relação aos resultados, pode-se dizer que o nível de erro apresentado pelos modelos, apesar de não serem demasiadamente grandes, podem fazer com que os mesmos não sejam suficientemente precisos, em um primeiro momento, para a efetiva adoção do modelo como um apoio à tomada de decisão por meio de sua capacidade preditiva. Tal mensuração do erro pode ser explicada pela existência de incertezas não facilmente representáveis em atributos no ambiente onde a organização se insere. Entretanto, tal comportamento não compromete o atingimento do OE3, cujo foco não é de encontrar a melhor performance possível, e sim validar a utilização de ML para solução do problema exposto, por meio da validação de um modelo que possa explicar os gastos de forma significativa. Além disso, as técnicas adicionais aplicadas para interpretação do modelo foram capazes de gerar percepções úteis para a tomada de decisão, como por exemplo, identificar quais atributos são os mais relevantes na composição da performance do modelo, bem como entender de maneira detalhada o relacionamento destes atributos com o gasto, por meio da análise da variabilidade do resultado final ao se implantar modificações nos valores dos atributos.

Finalmente, o quarto objetivo específico “OE4 - Avaliar se o modelo pode ser implantado de forma sistêmica dentro da organização, com o apoio de ferramentas tecnológicas de código aberto (software livre), de preferência” foi alcançado, inicialmente, por utilizar software livre em todas as etapas deste estudo, conforme premissas descritas na seção 3.4 e, posteriormente, ao longo do Capítulo 5, pela sugestão de aprimoramentos a serem considerados sobre um processo padrão de ETL para permitir que a organização se beneficie de forma continuada da abordagem desenvolvida.

6.1 Contribuições da Pesquisa

Seguem as principais contribuições advindas do presente estudo:

- Os resultados alcançados são capazes de produzir previsões que podem ser utilizadas como um critério adicional a ser levado em consideração ao longo do processo de tomada de decisão da organização. O conhecimento sobre os atributos mais relevantes para a performance obtida suporta que os modelos podem ser utilizados no dimensionamento dos gastos oriundos de decisões que envolvam, por exemplo, a movimentação urgente de tropas pelo território nacional, ou a ativação/desativação de instalações da organização, por serem capazes de gerar estimativas de redução ou acréscimo nos gastos com base em diversas características como quantidade de pessoal, atividade principal da instalação, nível hierárquico e estrutura organizacional.
- A abordagem conceitual foi desenvolvida com foco na reutilização e flexibilidade, de forma que seja facilmente aplicável no aprofundamento do problema tratado neste estudo, bem como em outros contextos, sem que haja a necessidade de profundo conhecimento prévio sobre o problema e com uma menor participação de especialistas na área de tecnologia e gestão. Dessa forma, uma parte do esforço deste trabalho objetivou viabilizar o aproveitamento futuro desta abordagem como uma ferramenta, cuja discussão de sua implementação é tratada em um capítulo próprio.
- O conhecimento dos atributos mais relevantes pode orientar futuros esforços de coleta de dados em direção a uma melhoria mais significativa na performance dos modelos. Além disso, outro potencial contributo da implementação da abordagem conceitual deste estudo é usufruir, mesmo que superficialmente, das vantagens de orientar a decisão com base em dados. Neste estágio inicial, a geração de percepções que sejam suficientemente úteis pode auxiliar a organização a superar as dificuldades expostas na seção 2.2.2, relacionadas ao tratamento do dado como um ativo organizacional essencial para formulação de sua estratégia. Quanto maior for a importância do dado dentro da organização, mais atributos podem ser mensurados e tratados com maior atenção, resultando em melhor performance final dos modelos treinados.
- Por fim, acredita-se que este estudo tenha contribuído para ampliar a discussão relacionada a inovações tecnológicas no setor público, uma vez constatada a existência de um relacionamento estreito entre as atividades governamentais e a aplicação de ferramentas

tecnológicas para os mais diversos fins, para obter resultados positivos para a sociedade de uma forma generalizada.

6.2 Limitações do Estudo

Os dados coletados foram considerados a maior limitação deste estudo. Manipular dados secundários que foram originalmente concebidos para atender outros propósitos que não se relacionam com este estudo se demonstrou uma tarefa complexa, uma vez que foram consultadas fontes, formas de acesso e de armazenamento completamente distintos. Além disso, após a coleta, nem todos os dados puderam ser integrados de forma a representar características de cada instalação da Marinha do Brasil, por fugir do propósito principal da sua existência em sua origem. Foi o caso dos dados coletados para compor os atributos FNMQ (metros quadrados de área construída) e FNPT (patrimônio), onde foi identificado que, em sua origem, os mesmos não são armazenados de forma individualizada para cada instalação e sim por meio de grandes aglomerados de instalações em uma determinada localização geográfica, sendo impossível integrá-los aos demais atributos.

Outra limitação foi a falta de pesquisas anteriores com as mesmas peculiaridades, onde uma única organização busca um modelo que a permita entender os gastos entre centenas de instalações com atividades bastante distintas entre si e, portanto, com demandas por recursos diferentes. A revisão da literatura apontou que poucos estudos se propõem a apoiar a decisão na gestão do orçamento na área de Defesa, e também não foi possível encontrar estudos que o fazem por categoria de despesa. Desta forma, foi preferível adotar uma abordagem bastante simples, com o propósito de validar a metodologia utilizada para o entendimento do problema ao invés de priorizar a aplicação e o refinamento de técnicas para maximizar a performance preditiva dos modelos encontrados.

6.3 Sugestões para Futuras Pesquisas

Seguem as sugestões para futuras pesquisas, idealizadas com base em ideias construídas ao longo do estudo realizado:

- A abordagem conceitual desenvolvida pode ser integrada a outros métodos úteis para apoio a decisão com o propósito de aprimorar a geração de informações úteis ao decisor. Um estudo poderia, por exemplo, aplicar *Data Envelopment Analysis* (DEA) de forma a obter uma medida de performance relativa de cada instalação, de forma a possibilitar a identificação de boas práticas das instalações com os melhores resultados. Além disso, ML poderia ser aplicado para estudar a performance relativa como um atributo-alvo do

aprendizado, para verificar se é possível aprender o que torna uma instalação da organização eficiente em seus gastos.

- Uma vez identificada a existência de incertezas não facilmente representáveis em atributos no ambiente onde a organização se insere, um estudo poderia ser realizado para inserir tais incertezas por meio de um processo estocástico para que, posteriormente, pudesse ser verificado se um modelo de ML é capaz de obter melhoria em sua performance relativa a fração da variabilidade não explicada anteriormente.
- O presente estudo pode ser repetido com foco na performance final dos modelos. Para tal, uma nova coleta de dados primários poderia ser efetuada para solucionar o problema de incompatibilidade de alguns tributos sofrido nesta pesquisa. Além disso, as próprias instalações poderiam fornecer outras características relevantes para serem utilizados como atributos adicionais. Dados de vários anos podem ser coletados para que seja possível estudar tendências e evoluções de uma série temporal. Por fim, diversas técnicas adicionais de ML poderiam ser aplicadas, e uma quantidade ainda maior de modelos poderiam ser treinados.
- Um outro estudo poderia se especializar na questão comportamental da apresentação dos resultados, de forma a sugerir maneiras que possibilitem que o decisor interaja com os modelos preditivos fruto do aprendizado de máquina, de forma a aproveitar todo o seu potencial. De uma forma geral, considerando os recentes avanços da inteligência artificial, sugerir o que deve ser feito para considerar as limitações cognitivas comuns a todos os humanos para usufruir do conhecimento gerado de forma autônoma, por computadores, com base em dados.

REFERÊNCIAS BIBLIOGRÁFICAS

- Abiodun, O. I., Jantan, A., Omolara, A. E., Dada, K. V., Mohamed, N. A. E., & Arshad, H. (2018). State-of-the-art in artificial neural network applications: A survey. *Heliyon*, *4*(11), e00938. <https://doi.org/10.1016/j.heliyon.2018.e00938>
- Agbozo, E., & Asamoah, B. K. (2019). Data-driven e-government: Exploring the socio-economic ramifications. *EJournal of EDemocracy and Open Government*, *11*(1), 81–90. <https://doi.org/10.29379/jedem.v11i1.510>
- Alptekin, A., & Levine, P. (2012). Military expenditure and economic growth: A meta-analysis. *European Journal of Political Economy*, *28*(4), 636–650. <https://doi.org/10.1016/j.ejpoleco.2012.07.002>
- Anderson, C. (2015). *Creating a data-driven organization: Practical advice from the trenches*. Sebastopol: O'Reilly Media, Inc.
- Barnett, V., & Lewis, T. (1978). *Outliers in statistical data* (1st ed.). Belfast: Wiley.
- Bekkers, V., & Homburg, V. (2005). *The Information Ecology of E-Government: E-Government as Institutional and Technological Innovation in Public Administration* (Vol. 9). Amsterdam: IOS press.
- Bilal, M., & Oyedele, L. O. (2020). Guidelines for applied machine learning in construction industry—A case of profit margins estimation. *Advanced Engineering Informatics*, *43*, 101013. <https://doi.org/10.1016/j.aei.2019.101013>
- Boehmke, B., & Greenwell, B. M. (2019). *Hands-On Machine Learning with R*. Boca Raton, Florida: CRC Press.
- Boueri, R., Rocha, F., & Rodopoulos, F. (2015). *Avaliação da Qualidade do Gasto Público e Mensuração da Eficiência* (R. Boueri, F. Rocha, & F. Rodopoulos, Eds.). Brasília: Secretaria do Tesouro Nacional.
- Bouman, B. A. M., & van Laar, H. H. (2006). Description and evaluation of the rice growth model ORYZA2000 under nitrogen-limited conditions. *Agricultural Systems*, *87*(3), 249–273. <https://doi.org/https://doi.org/10.1016/j.agsy.2004.09.011>
- Box, G. E. P., & Cox, D. R. (1964). An Analysis of Transformations. *Journal of the Royal Statistical Society: Series B (Methodological)*, *26*(2), 211–243. <https://doi.org/10.1111/j.2517-6161.1964.tb00553.x>
- Brasil. Presidência da República. (2011, May). Lei 12.527, de 18 de novembro de 2011. Regula o

- acesso a informações e dá outras providências. *Diário Oficial Da União*. Retrieved from http://www.planalto.gov.br/ccivil_03/_ato2011-2014/2011/lei/l12527.htm
- Breiman, L. (2001). Random Forests. *Machine Learning*, 45(1), 5–32. <https://doi.org/10.1023/A:1010933404324>
- Brynjolfsson, E., Hitt, L., & Kim, H. (2011). Strength in numbers: How does data-driven decision-making affect firm performance? *International Conference on Information Systems 2011, ICIS 2011*, 1, 541–558.
- Burrell, G., & Morgan, G. (1979). *Sociological Paradigms and Organisational Analysis: Elements of the Sociology of Corporate Life*. Retrieved from https://books.google.pt/books?id=Q_62AAAAIAAJ
- Caruso, R., & Francesco, A. (2012). Country Survey: Military Expenditure And Its Impact On Productivity In Italy, 1988-2008. *Defence and Peace Economics*, 23(5), 471–484. <https://doi.org/10.1080/10242694.2011.608964>
- Chambers, R., Hentges, A., & Zhao, X. (2004). Robust automatic methods for outlier and error detection. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 167(2), 323–339. <https://doi.org/10.1111/j.1467-985X.2004.00748.x>
- Chen, P. P.-S. (1976). The Entity-Relationship Model—toward a Unified View of Data. *ACM Trans. Database Syst.*, 1(1), 9–36. <https://doi.org/10.1145/320434.320440>
- Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794. <https://doi.org/10.1145/2939672.2939785>
- Chlingaryan, A., Sukkarieh, S., & Whelan, B. (2018). Machine learning approaches for crop yield prediction and nitrogen status estimation in precision agriculture: A review. *COMPUTERS AND ELECTRONICS IN AGRICULTURE*, 151, 61–69. <https://doi.org/10.1016/j.compag.2018.05.012>
- Christodoulou, P., Decker, S., Douka, A.-V., Komopoulou, C., Peristeras, V., Sgagia, S., ... Vardouniotis, D. (2018). *Data Makes the Public Sector Go Round BT - Electronic Government* (P. Parycek, O. Glassey, M. Janssen, H. J. Scholl, E. Tambouris, E. Kalampokis, & S. Virkar, Eds.). Cham: Springer International Publishing.
- Davenport, T., & Harris, J. (2017). *Competing on analytics: Updated, with a new introduction: The new science of winning*. Boston: Harvard Business Press.

- De Rezende, L. B., & Blackwell, P. (2019). The Brazilian National Defence Strategy: Defence Expenditure Choices and Military Power. *Defence and Peace Economics*, 1–16. <https://doi.org/10.1080/10242694.2019.1588030>
- de Vries, H., Bekkers, V., & Tummers, L. (2016). Innovation in the Public Sector: A Systematic Review and Future Research Agenda. *Public Administration*, 94(1), 146–166. <https://doi.org/10.1111/padm.12209>
- Du, M., Liu, N., & Hu, X. (2019). Techniques for Interpretable Machine Learning. *Commun. ACM*, 63(1), 68–77. <https://doi.org/10.1145/3359786>
- Edwards, R. E., New, J., & Parker, L. E. (2012). Predicting future hourly residential electrical consumption: A machine learning case study. *Energy and Buildings*, 49, 591–603. <https://doi.org/10.1016/j.enbuild.2012.03.010>
- El Naqa, I., & Murphy, M. J. (2015). *What Is Machine Learning? BT - Machine Learning in Radiation Oncology: Theory and Applications* (I. El Naqa, R. Li, & M. J. Murphy, Eds.). https://doi.org/10.1007/978-3-319-18305-3_1
- Feaver, P. D. (1999). Civil-military relations. *ANNUAL REVIEW OF POLITICAL SCIENCE*, 2, 211–241. <https://doi.org/10.1146/annurev.polisci.2.1.211>
- Feilhauer, H., Asner, G. P., Martin, R. E., & Schmidtlein, S. (2010). Brightness-normalized Partial Least Squares Regression for hyperspectral data. *Journal of Quantitative Spectroscopy and Radiative Transfer*, 111(12), 1947–1957. <https://doi.org/https://doi.org/10.1016/j.jqsrt.2010.03.007>
- Ferreira, F. G. B. de C., & Oliveira, C. L. (2017). O orçamento público no Estado constitucional democrático e a deficiência crônica na gestão das finanças públicas no Brasil. *Seqüência: Estudos Jurídicos e Políticos*, 38(76), 183. <https://doi.org/10.5007/2177-7055.2017v38n76p183>
- Fisher, A., Rudin, C., & Dominici, F. (2019). All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. *Journal of Machine Learning Research*, 20(177), 1–81.
- Friedman, J. H. (2001). Greedy Function Approximation: A Gradient Boosting Machine. *The Annals of Statistics*, 29(5), 1189–1232. Retrieved from <http://www.jstor.org/stable/2699986>
- Ghalekhondabi, I., Ardjmand, E., Young, W. A., & Weckman, G. R. (2017). Water demand forecasting:

- review of soft computing methods. *Environmental Monitoring and Assessment*, 189(7), 313. <https://doi.org/10.1007/s10661-017-6030-3>
- Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT press.
- Hawkins, D. M. (2004). The Problem of Overfitting. *Journal of Chemical Information and Computer Sciences*, 44(1), 1–12. <https://doi.org/10.1021/ci0342472>
- Hodge, V., & Austin, J. (2004). A Survey of Outlier Detection Methodologies. *Artificial Intelligence Review*, 22(2), 85–126. <https://doi.org/10.1023/B:AIRE.0000045502.10941.a9>
- Hong, H., Pradhan, B., Bui, D. T., Xu, C., Youssef, A. M., & Chen, W. (2017). Comparison of four kernel functions used in support vector machines for landslide susceptibility mapping: a case study at Suichuan area (China). *Geomatics, Natural Hazards and Risk*, 8(2), 544–569. <https://doi.org/10.1080/19475705.2016.1250112>
- Hou, D. (2018). The determinants of military expenditure in Asia and Oceania, 1992-2016: A dynamic panel analysis. *Peace Economics, Peace Science and Public Policy*, Vol. 24. <https://doi.org/10.1515/peps-2018-0004>
- Huang, J., Li, Y.-F., & Xie, M. (2015). An empirical analysis of data preprocessing for machine learning-based software cost estimation. *Information and Software Technology*, 67, 108–127. <https://doi.org/https://doi.org/10.1016/j.infsof.2015.07.004>
- Instituto Nacional de Meteorologia - INMET. (2019). BDMEP - Banco de Dados Meteorológicos para Ensino e Pesquisa. Retrieved July 19, 2019, from <http://www.inmet.gov.br/portal/index.php?r=bdmep/bdmep>
- Instituto Brasileiro de Geografia e Estatística. (2019). Cidades e Estados: Rondônia. Retrieved January 22, 2020, from <https://www.ibge.gov.br/pt/cidades-e-estados.html>
- Jang, H. (2019). A decision support framework for robust R&D budget allocation using machine learning and optimization. *Decision Support Systems*, 121, 1–12. <https://doi.org/10.1016/j.dss.2019.03.010>
- Janssen, M., & Estevez, E. (2013). Lean government and platform-based governance-Doing more with less. *Government Information Quarterly*, 30(SUPPL. 1), S1–S8. <https://doi.org/10.1016/j.giq.2012.11.003>
- Janssen, M., Konopnicki, D., Snowdon, J. L., & Ojo, A. (2017). Driving public sector innovation using

- big and open linked data (BOLD). *Information Systems Frontiers*, 19(2), 189–195.
<https://doi.org/10.1007/s10796-017-9746-2>
- Kartal, H., Oztekin, A., Gunasekaran, A., & Cebi, F. (2016). An integrated decision analytic framework of machine learning with multi-criteria decision making for multi-attribute inventory classification. *Computers and Industrial Engineering*, 101, 599–613.
<https://doi.org/10.1016/j.cie.2016.06.004>
- Kersting, K. (2018). Machine Learning and Artificial Intelligence: Two Fellow Travelers on the Quest for Intelligent Behavior in Machines . *Frontiers in Big Data* , Vol. 1, p. 6. Retrieved from <https://www.frontiersin.org/article/10.3389/fdata.2018.00006>
- Kiron, D. (2017). Lessons from Becoming a Data-Driven Organization. *MIT Sloan Management Review*, 58(2).
- Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. *Ijcai*, 14(2), 1137–1145. Montreal, Canada.
- Kohavi, R., & John, G. H. (1997). Wrappers for feature subset selection. *Artificial Intelligence*, 97(1–2), 273–324.
- Kuhn, M. (2008). Building Predictive Models in R Using the caret Package. *Journal of Statistical Software; Vol 1, Issue 5 (2008)* . Retrieved from <https://www.jstatsoft.org/v028/i05>
- Kuhn, M., & Johnson, K. (2013). *Applied predictive modeling* (Vol. 26). New York: Springer.
- Kumar, V., Chattaraman, V., Neghina, C., Skiera, B., Aksoy, L., Buoye, A., & Henseler, J. (2013). Data-driven services marketing in a connected world. *Journal of Service Management*, 24(3), 330–352. <https://doi.org/10.1108/09564231311327021>
- Kursa, M. B., & Rudnicki, W. R. (2010). Feature Selection with the {Boruta} Package. *Journal of Statistical Software*, 36(11), 1–13. Retrieved from <http://www.jstatsoft.org/v36/i11/>
- Kvålseth, T. O. (1985). Cautionary note about R 2. *The American Statistician*, 39(4), 279–285.
- LaValle, S., Lesser, E., Shockley, R., Hopkins, M. S., & Kruschwitz, N. (2011). Big Data, Analytics and the Path From Insights to Value. *MIT Sloan Management Review*, 52(2), 21–32. Retrieved from <https://search.proquest.com/docview/845235605?accountid=39260>
- Lee, K. M., Yoo, J., Kim, S. W., Lee, J. H., & Hong, J. (2019). Autonomic machine learning platform. *International Journal of Information Management*, 49, 491–501.

<https://doi.org/10.1016/j.ijinfomgt.2019.07.003>

- Liu, F. T., Ting, K. M., & Zhou, Z.-H. (2008). Isolation Forest. *2008 Eighth IEEE International Conference on Data Mining*, 413–422. <https://doi.org/10.1109/ICDM.2008.17>
- Marinha do Brasil. (2014). *Normas para a gestão do Sistema do Plano Diretor (SGM-401 Rev. 1)*. Brasil.
- Marinha do Brasil. (2019). Estrutura Organizacional. Retrieved January 12, 2020, from <https://www.marinha.mil.br/estrutura-organizacional>
- Markham, I. S., & Rakes, T. R. (1998). The effect of sample size and variability of data on the comparative performance of artificial neural networks and regression. *Computers & Operations Research*, *25*(4), 251–263. [https://doi.org/https://doi.org/10.1016/S0305-0548\(97\)00074-9](https://doi.org/https://doi.org/10.1016/S0305-0548(97)00074-9)
- Millard, J. (2018). Open governance systems: Doing more with more. *Government Information Quarterly*, *35*(4), S77–S87. <https://doi.org/10.1016/j.giq.2015.08.003>
- Ministério da Defesa. (2019a). Ações subsidiárias realizadas pela Marinha do Brasil. Retrieved January 12, 2020, from <https://www.defesa.gov.br/programas-sociais/acoes-subsidiarias/acoes-subsidiarias-realizadas-pela-marinha-do-brasil>
- Ministério da Defesa. (2019b). Marinha do Brasil. Retrieved January 12, 2020, from <https://www.defesa.gov.br/forcas-armadas/marinha-do-brasil>
- Morrel-Samuels, P., Francis, E., & Shucard, S. (2009). Merged Datasets: An Analytic Tool for Evidence-Based Management. *California Management Review*, *52*(1), 120–139. <https://doi.org/10.1525/cm.2009.52.1.120>
- Myrtveit, I., Stensrud, E., & Olsson, U. H. (2001). Analyzing data sets with missing data: An empirical evaluation of imputation methods and likelihood-based methods. *IEEE Transactions on Software Engineering*, *27*(11), 999–1013. <https://doi.org/10.1109/32.965340>
- Pallonetto, F., De Rosa, M., Milano, F., & Finn, D. P. (2019). Demand response algorithms for smart-grid ready residential buildings using machine learning models. *Applied Energy*, *239*, 1265–1282. <https://doi.org/10.1016/j.apenergy.2019.02.020>
- Provost, F., & Fawcett, T. (2013). Data Science for Business: What You Need to Know About Data Mining and Data-Analytic Thinking. In *Sebastopol, Calif. : O'Reilly*. “ O'Reilly Media, Inc.”
- Quinlan, J. R. (1986). Induction of decision trees. *Machine Learning*, *1*(1), 81–106.

- R Core Team. (2017). *R: A Language and Environment for Statistical Computing*. Retrieved from <https://www.r-project.org/>
- Rath, M., Pattanayak, B. K., & Pati, B. (2016). Energy Efficient MANET Protocol Using Cross Layer Design for Military Applications. *Defence Science Journal*, *66*(2), 146–150. <https://doi.org/10.14429/dsj.66.9705>
- Robinson, C., Dilkina, B., Hubbs, J., Zhang, W., Guhathakurta, S., Brown, M. A., & Pendyala, R. M. (2017). Machine learning approaches for estimating commercial building energy consumption. *Applied Energy*, *208*, 889–904. <https://doi.org/10.1016/j.apenergy.2017.09.060>
- Saunders, M. N. K., Lewis, P., & Thornhill, A. (2016). *Research methods for business students* (Seventh). Harlow: Pearson Education Limited.
- Secretaria do Orçamento Federal. (2019). *Manual Técnico de Orçamento*. Brasília, DF. Brasil.
- Secretaria do Tesouro Nacional. (2019). SIAFI - STN. Retrieved March 1, 2020, from <http://www.stn.fazenda.gov.br/SIAFI>
- Singh, K. (2007). *Quantitative Social Research Methods*. <https://doi.org/10.4135/9789351507741>
- SIPRI - The World Bank. (2018). Military expenditure (% of GDP) | Data. Retrieved January 28, 2020, from <https://data.worldbank.org/indicator/MS.MIL.XPND.GD.ZS>
- Smola, A. J., & Schölkopf, B. (2004). A tutorial on support vector regression. *Statistics and Computing*, *14*(3), 199–222. <https://doi.org/10.1023/B:STCO.0000035301.49549.88>
- Stablein, R. (1999). *Data in organization studies*. London: Sage Publications.
- Stockholm International Peace Research Institute - SIPRI. (2018). Sources and methods | SIPRI. Retrieved January 28, 2020, from <https://www.sipri.org/databases/milex/sources-and-methods>
- Sudhakar, I., Madhusudhan Reddy, G., & Srinivasa Rao, K. (2016). Ballistic behavior of boron carbide reinforced AA7075 aluminium alloy using friction stir processing – An experimental study and analytical approach. *Defence Technology*, *12*(1), 25–31. <https://doi.org/10.1016/j.dt.2015.04.005>
- Svensen, N. G., Kalita, P. K., & Gebhart, D. L. (2005). Environmental risk reduction and combat readiness enhancement of military training lands through range design and maintenance. *2005 ASAE Annual International Meeting*.
- Szuster, N., Cardoso, R. L., Szuster, F. R., Szuster, F. R., & Szuster, F. R. (2009). Contabilidade Geral:

introdução à contabilidade societária. In *São Paulo: Atlas*.

- Touzani, S., Granderson, J., & Fernandes, S. (2018). Gradient boosting machine for modeling the energy consumption of commercial buildings. *Energy and Buildings*, *158*, 1533–1543. <https://doi.org/https://doi.org/10.1016/j.enbuild.2017.11.039>
- Valdes, G., Simone, C. B., Chen, J., Lin, A., Yom, S. S., Pattison, A. J., ... Solberg, T. D. (2017). Clinical decision support of radiotherapy treatment planning: A data-driven machine learning strategy for patient-specific dosimetric decision making. *Radiotherapy and Oncology*, *125*(3), 392–397. <https://doi.org/10.1016/j.radonc.2017.10.014>
- van Buuren, S., & Groothuis-Oudshoorn, K. (2011). {mice}: Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software*, *45*(3), 1–67. Retrieved from <https://www.jstatsoft.org/v45/i03/>
- Van der Laan, M. J., Polley, E. C., & Hubbard, A. E. (2007). Super learner. *Statistical Applications in Genetics and Molecular Biology*, *6*(1). <https://doi.org/10.2202/1544-6115.1309>
- Xiao, Y., Wu, J., Lin, Z., & Zhao, X. (2018). A deep learning-based multi-model ensemble method for cancer prediction. *Computer Methods and Programs in Biomedicine*, *153*, 1–9. <https://doi.org/10.1016/j.cmpb.2017.09.005>
- Yaseen, Z. M., Deo, R. C., Hilal, A., Abd, A. M., Bueno, L. C., Salcedo-Sanz, S., & Nehdi, M. L. (2018). Predicting compressive strength of lightweight foamed concrete using extreme learning machine model. *Advances in Engineering Software*, *115*, 112–125. <https://doi.org/10.1016/j.advengsoft.2017.09.004>
- Ye, H., Liang, L., Li, G. Y., Kim, J., Lu, L., & Wu, M. (2018). Machine Learning for Vehicular Networks: Recent Advances and Application Examples. *IEEE Vehicular Technology Magazine*, *13*(2), 94–101. <https://doi.org/10.1109/MVT.2018.2811185>
- Zhang, S., Zhang, C., & Yang, Q. (2003). Data preparation for data mining. *Applied Artificial Intelligence*, *17*(5–6), 375–381. <https://doi.org/10.1080/713827180>

APÊNDICE I – Relação de despesas enquadradas em Funcionamento (FC)

Funcionamento – Relação de ND	
____3004 – Gás Engarrafado	____3947 – Serviços de Comunicação em Geral
____3625 – Serviços de Limpeza e Conservação (PF)	____3958 – Serviços de Telecomunicações
____3943 – Serviços de Energia Elétrica	____3978 – Limpeza e Conservação (PJ)
____3944 – Serviços de Água e Esgoto	____3997 – Comunicação de Dados
____3945 – Serviços de Gás	____4014 – Pacote de Transmissão de Dados
____4722 – Contribuição de Iluminação Pública	____4710 – Taxas

APÊNDICE II – Relação de despesas enquadradas em Manutenção (MN)

Manutenção – Relação de ND	
____3001 – Combustíveis e Lubrificantes automotivos	____3622 – Manutenção e Conservação de bens imóveis (PF)
____3016 – Material de Expediente	____3654 – Manutenção e Conservação de Equipamentos de Processamento de Dados (PF)
____3017 – Material de TIC (consumo)	____4000 – Serviços de Tecnologia da Informação e Comunicação (TIC)
____5235 – Material de TIC (permanente)	____3669 – Seguros em Geral (PF)
____3022 – Material de Limpeza e Produção de Higienização	____3689 – Manutenção de Repartições, serviço exterior (PF)
____3024 – Material para manutenção de bens imóveis	____3912 – Locação de Máquinas e Equipamentos
____3025 – Material para manutenção de bens móveis (exceto veículos)	____3916 – Manutenção e Conservação de bens imóveis (PJ)
____3026 – Material Elétrico e Eletrônico	____3917 – Manutenção e Conservação de máquinas e equipamentos
____3039 – Material para Manutenção de Veículos	____3919 – Manutenção e Conservação de veículos
____4006 – Locação de Softwares	____3920 – Manutenção e Conservação de bens móveis de outras naturezas
____3089 – Material de Consumo, Repartições no Exterior	____3969 – Seguros em Geral (PJ)
____3099 – Outros materiais de consumo	____3983 – Cópias e Reprodução de Documentos
____3618 – Manutenção e Conservação de Equipamentos (PF)	____3989 – Manutenção de Repartições, serviço exterior (PJ)
____3620 – Manutenção e Conservação de Veículos (PF)	____3990 – Serviço de Publicidade Legal
____3621 – Manutenção e Conservação de bens móveis de outras naturezas	____3995 – Manutenção e Conservação de Equipamentos de Processamento de Dados (PJ)

