



Universidade Federal de Uberlândia
Faculdade de Engenharia Elétrica
Graduação em Engenharia de Controle e Automação

BRUNA CORRÊA DE MORAIS

**SISTEMA INTEGRADO DE ANÁLISE E PREDIÇÃO DE INDICADORES DE
DESEMPENHO DE UM PROCESSO INDUSTRIAL**

Uberlândia
2019

BRUNA CORRÊA DE MORAIS

**SISTEMA INTEGRADO DE ANÁLISE E PREDIÇÃO DE INDICADORES DE
DESEMPENHO DE UM PROCESSO INDUSTRIAL**

Trabalho apresentado como requisito parcial de avaliação na disciplina Trabalho de Conclusão de Curso do Curso de Engenharia de Controle e Automação da Universidade Federal de Uberlândia.

Orientador: Dr. Renato Ferreira Fernandes Jr.

Assinatura do orientador

Uberlândia

2019

BRUNA CORRÊA DE MORAIS

**SISTEMA INTEGRADO DE ANÁLISE E PREDIÇÃO DE INDICADORES DE
DESEMPENHO DE UM PROCESSO INDUSTRIAL**

Trabalho apresentado como requisito parcial de avaliação na disciplina Trabalho de Conclusão de Curso do Curso de Engenharia de Controle e Automação da Universidade Federal de Uberlândia.

Banca de Avaliação:

Prof. Dr. Renato Ferreira Fernandes Jr
Orientador

Prof. Dr. Josué Silva de Moraes
Membro

Prof. Dr. Renato Santos Carrijo
Membro

Uberlândia (MG), 11 de dezembro de 2019.

Morais, Bruna Corrêa

SISTEMA INTEGRADO DE ANÁLISE E PREDIÇÃO DE INDICADORES DE DESEMPENHO DE UM PROCESSO INDUSTRIAL. **Bruna Corrêa de Moraes. – UBERLÂNDIA, 2019** - 91p, 30 cm.

Orientador: Prof. Dr. Renato Ferreira Fernandes Júnior

Trabalho de Conclusão de Curso – Universidade Federal de Uberlândia – UFU Faculdade de Engenharia de Elétrica. 2019.

1. Business Intelligence 2. Monitoramento 3. Análises 4. Processo Industrial 5. Aprendizado de Máquinas 6. Predição

7. I. Prof. Dr. Renato Ferreira Fernandes Júnior II. Universidade Federal de Uberlândia III. Faculdade de Engenharia Elétrica IV. Engenharia de Controle e Automação

Aos meus pais, irmãos e noivo que, com tanto carinho e apoio, não mediram esforços para que eu chegasse ao final dessa jornada.

AGRADECIMENTOS

Agradeço à Deus pela vida e por todas suas bênçãos que me foram concedidas.

Aos meus pais, Lionilda e Marcio, pelos ensinamentos e apoio incondicional durante toda a minha vida.

Aos meus irmãos, gratidão pelo incentivo e inspiração que sempre representaram para mim.

Ao meu noivo Rubens, por todo o carinho, compreensão e companheirismo.

Ao professor orientador Renato Fernandes, pela paciência, disponibilidade e instrução no decorrer do projeto, além dos conhecimentos transmitidos ao longo do curso.

RESUMO

Este trabalho apresenta a construção de um sistema integrado de análise e predição de indicadores referentes a um processo industrial, o qual tem como objetivo suportar a tomada de decisão no ambiente empresarial, baseando-se em dados históricos. O sistema desenvolvido utilizou-se dos conceitos de Business Intelligence para a construção de dashboards interativos, e de Aprendizado de Máquinas para modelagem de algoritmos preditivos capazes de antecipar possíveis desvios no produto acabado. Para isso, foi utilizado um sistema real utilizando tecnologias OPC, SQL Server e Power BI para monitoramento e análises. Para o aprendizado de Máquinas, foi utilizado o modelo de classificação K-NN, através da linguagem Python. A validação da ferramenta criada é apresentada, além dos resultados obtidos de sua utilização no ambiente industrial proposto.

Palavras Chave: Business Intelligence, monitoramento, análises, processo industrial, Aprendizado de Máquinas, predição.

ABSTRACT

This paper presents the construction of an integrated system for analysis and prediction of indicators related to an industrial process, which aims to support decision making in the business environment, based on historical data. The developed system uses the concepts of Business Intelligence for the construction of interactive dashboards, and Machine Learning for modeling predictive algorithms capable of anticipating possible deviations in the finished product. For this, a real system was built using OPC, SQL Server and Power BI technologies for monitoring and analysis. For machine learning, we used the K-NN classification model, using the Python language. The validation of the created tool is presented, besides the results obtained from its use in the proposed industrial environment.

Keywords: *Business Intelligence, monitoring, analysis, industrial process, Machine Learning, prediction.*

LISTAS DE ABREVIATURAS E SIGLAS

BI: Business Intelligence

BPM: Business Performance Manager

CRUD: Create, Read, Update, Delete

DAX: Data Analysis Expressions

DM: Data Mart

DW: Data Warehouse

ETL: Extract, Transform, Load

FC: Function

IBM: International Business Machine

IGS: Industrial Gateway Server

IoT: Internet of Things

K-NN: K-Nearst Neighbors

ML: Machine Learning

OLAP: Online Analytical Processing

OLE: Object Linking and Embedding

OLTP: Online Transaction Processing

OPC: OLE for Process Control

PDCA: Plan, Do, Check, Act

PLC: Programmable Logic Controller (CLP – Controlador Lógico Programável)

PNL: Processamento de Linguagem Natural

TCP: Transmission Control Protocol

LISTA DE FIGURAS

Figura 1: Cadeia de Valores do BI	19
Figura 2: Ciclo da Informação	20
Figura 3: Cubo da Informação.....	22
Figura 4: Modelos OLTP e OLAP.....	24
Figura 5: Processo de ETL.....	25
Figura 6:Relação de número de decisões versus o valor das decisões para o negócio.....	28
Figura 7: Quadrante Mágico das soluções de BI de mercado.....	30
Figura 8: Microsoft Power BI	31
Figura 9: Tableau	33
Figura 10: Exemplo de Aprendizado Supervisionado.....	38
Figura 11: Exemplo de Aprendizado Não Supervisionado.	39
Figura 12: Exemplo de Aprendizado Semisupervisionado.	40
Figura 13: Exemplo de Aprendizado por Reforço.....	41
Figura 14: Aprendizado Online.....	42
Figura 15: Comparativo entre Aprendizado Online e por Batelada.	42
Figura 16: Aprendizado baseado em instâncias.....	43
Figura 17: Aprendizado baseado em modelo.....	44
Figura 18: Esquema geral de um algoritmo tradicional.	44
Figura 19: Esquema geral de um algoritmo de Machine Learning.	45
Figura 20: Modelos de Aprendizado de Máquinas.	45
Figura 21: Comparativo entre modelos de classificação e regressão.	46
Figura 22: Aprendizagem preditiva.....	47
Figura 23: Exemplo de funcionamento do modelo K-NN.	48
Figura 24: Matriz de Confusão.	51
Figura 25: Esquema geral.	52
Figura 26-: PLCs S7-300 e S7-400 Siemens	53
Figura 27: Ambiente do Step 7.....	54
Figura 28: Exemplo de lógica de análise de equipamentos em funcionamento.	55
Figura 29: Exemplo de lógica para mover variáveis no Step 7.....	55
Figura 30: RunTime do IGS.....	57
Figura 31: Configuração de dispositivos no IGS.	58
Figura 32: Configuração de uma variável do tipo Word no IGS.	58
Figura 33: Configuração de uma variável booleana no IGS.....	59
Figura 34: Ambiente do OPC Data Logger.....	60
Figura 35: Configuração de coleta com compactação de dados.....	61
Figura 36: Conexão do OPC Data Logger com tabelas do SQL Server.....	62
Figura 37: Novo Banco de Dados no SQL.	63
Figura 38: Exemplo de modelagem utilizada no Banco de Dados.	64
Figura 39: Nova tabela de dados no SQL.	64
Figura 40: Operações básicas no Query do SQL.....	65
Figura 41: Configuração de Tarefas agendadas no SQL Agent.....	66
Figura 42: Conexão do Power BI com a fonte de dados.	67
Figura 43: Configurações de modo de conexão do SQL com Power BI	67

Figura 44: Editor de Consultas do Power BI.....	68
Figura 45: Editor Avançado do Power BI.	69
Figura 46: Guia de dados do Power BI.	70
Figura 47: Exemplo de uma medida criada utilizando DAX.	70
Figura 48: Relacionamento entre tabelas na guia de Modelo do Power BI.....	71
Figura 49: Exemplo de relatório criado no Power BI.	72
Figura 50: Exemplo de análises criadas no Power BI.	73
Figura 51: Ambiente do Anaconda Navigator.....	76
Figura 52: Histograma dos dados de Qualidade.	77
Figura 53: Visão geral dos parâmetros de um dos conjuntos de dados.....	77
Figura 54: Parâmetros do conjunto dos dados de processo.	78
Figura 55: Pré-processamento dos dados de processo.	79
Figura 56: Aplicação do método K-NN.	79
Figura 57: Relatório criado para análise de curvas no Power BI.....	83
Figura 58: Análises no Power BI.	84
Figura 59: Matriz de Confusão dos Dados de Qualidade (didático).	86
Figura 60: Matriz de Confusão dos Dados do processo real.....	86

LISTA DE TABELAS

Tabela 1: Comparação entre dado, informação e conhecimento	20
Tabela 2: Comparação entre Data Marts e Data Warehouse	23
Tabela 3: Comparação entre OLTP e OLAP	24
Tabela 4: Características do BI em seus diferentes níveis.....	28
Tabela 5: Power BI x Tableau	33
Tabela 6: Relação da quantidade de PLCs e variáveis coletadas por área.	56
Tabela 7: Resultados dos testes de predição.....	85

SUMÁRIO

1	INTRODUÇÃO	13
2	OBJETIVOS	14
3	FUNDAMENTOS TEÓRICOS	15
3.1	BUSINESS INTELLIGENCE	15
3.1.1	<i>Conceitos</i>	15
3.1.2	<i>Histórico do Business Intelligence</i>	16
3.1.3	<i>Business Intelligence e a Cadeia de Valor</i>	18
3.1.4	<i>Ciclo da Informação e Arquitetura do Business Intelligence</i>	19
3.1.5	<i>Aplicações de BI nas organizações</i>	27
3.1.6	<i>Self-Service BI</i>	29
3.1.7	<i>Soluções de BI de mercado</i>	29
3.1.8	<i>Futuro do BI</i>	33
3.2	MACHINE LEARNING	35
3.2.1	<i>Conceitos</i>	35
3.2.2	<i>Histórico do Aprendizado de Máquinas</i>	36
3.2.3	<i>Tipos de Aprendizado de Máquinas</i>	37
3.2.4	<i>Composição básica de um algoritmo de Aprendizado de Máquinas</i>	44
3.2.5	<i>Modelos de Aprendizado de Máquinas</i>	45
3.2.6	<i>Desafios na Implementação de Algoritmos de Aprendizado de Máquinas</i>	48
3.2.7	<i>Validações e Testes</i>	50
4	DESENVOLVIMENTO	52
4.1	COLETA DE DADOS	53
4.1.1	<i>Dados de Processo</i>	53
4.1.2	<i>OPC Server – IGS</i>	56
4.1.3	<i>OPC Client – Advanced OPC Data Logger</i>	59
4.2	BANCO DE DADOS – SQL SERVER	62
4.3	DASHBOARDS DE MONITORAMENTO E ANÁLISES NO POWER BI	66
4.4	PREDIÇÕES DE DESVIOS	73
4.5	METODOLOGIA DE VALIDAÇÃO	80
5	RESULTADOS	81
5.1	VALIDAÇÃO DA COLETA E BANCO DE DADOS	81
5.2	ESTUDO DE CASO DO PASSO A PASSO DE DESENVOLVIMENTO E RESULTADOS OBTIDOS COM O USO DA FERRAMENTA	81
5.3	RESULTADOS DOS MODELOS DE PREDIÇÃO	85
6	CONCLUSÃO	87
7	REFERÊNCIAS BIBLIOGRÁFICAS	88

1 INTRODUÇÃO

A corrida por vantagem competitiva no mundo organizacional é uma realidade cada vez mais constante entre as empresas. No setor industrial, mais especificamente, a necessidade por entregar cada vez mais produtividade com menor custo exige dos executivos assertividade em suas tomadas de decisão. Isso implica em reduzir as paradas operacionais, que impactam em indicadores de desempenho, e constante melhoria na qualidade do produto, em busca de diferenciação de mercado. Desta maneira, decisões pautadas em informações capazes de gerar valor para a organização são necessárias. Para consolidação dessa informação, utiliza-se dados históricos de processos operacionais. Um importante conjunto de técnicas aplicado para esta área é o Business Intelligence, o qual vem assumindo importante papel para transformação da eficiência na tomada de decisão nas organizações.

Além da disponibilização da informação proveniente de dados extraídos da empresa, é preciso o reconhecimento de padrões e a predição de informações referentes ao processo e à qualidade do produto acabado, para que ações sejam tomadas antes de os problemas acontecerem. Isto é, uma atuação preditiva, baseada em informações passadas. É neste contexto que se utiliza o Aprendizado de Máquinas para construção de algoritmos capazes de fornecer tais análises.

Apesar do grande avanço que estas duas vertentes têm alcançado nos últimos anos, observa-se ainda pouca utilização dos enormes benefícios que podem ser agregados ao ambiente industrial. Desta maneira, o trabalho propõe-se a construir um sistema integrado que tem como objetivo disponibilizar análises e predições referentes à indicadores de desempenho de um processo industrial, com o objetivo de auxiliar nas tomadas de decisão da organização em questão.

2 OBJETIVOS

O presente trabalho apresenta a criação de uma ferramenta de monitoramento de informações de controle de processo de uma área industrial, cujo objetivo é fornecer análises complexas de dados referentes às variáveis de produção, oferecendo previsibilidade com precisão e acurácia, auxiliando na tomada de decisão da organização. Para isso, o sistema deverá realizar a coleta e armazenamento dos dados, disponibilizar as informações em relatórios visuais e gerar uma predição de indicadores de desempenho, baseado em fatos passados.

Desta forma, este trabalho foi dividido em seis capítulos. O primeiro capítulo foi dado uma breve introdução sobre o tema e apresentado os objetivos. No capítulo 2, são abordados os conceitos e fundamentações referentes ao Business Intelligence e às técnicas de predição de Aprendizados de Máquinas. No capítulo 3 é contemplado o desenvolvimento da ferramenta proposta, especificando cada uma das etapas para sua construção. O Capítulo 4 reserva-se aos resultados e validações do projeto em cada um de seus estágios. Por fim, o capítulo 5 apresenta as conclusões obtidas da produção do trabalho e sugestões de trabalhos futuros.

3 FUNDAMENTOS TEÓRICOS

Esse capítulo destina-se à revisão bibliográfica fundamental para a compreensão desse trabalho, o qual abordará os temas: Business Intelligence e Machine Learning.

3.1 Business Intelligence

3.1.1 Conceitos

Business Intelligence (BI), ou em português, Inteligência de Negócios, é um termo macro que engloba um conjunto de técnicas, conceitos e métodos voltados para apoiar a tomada de decisão de uma organização. Seu conceito tem diferentes abordagens para diferentes autores, uma vez que “Business Intelligence” se tornou uma expressão livre de conteúdo e utilizada amplamente para contextos diversos. Vercellis (2009) define o BI como apoio de modelos matemáticos e metodologias de análise que utiliza dados para tomadas de decisão complexas. Já Luhn (1958) reforça a auto-abstração e auto-codificação de documentos diferentes para pontos da organização.

Pode-se definir os objetivos do BI, segundo Turban (2009):

“Os principais objetivos do Business Intelligence são permitir o acesso interativo aos dados (às vezes em tempo real), proporcionar a manipulação desses dados e fornecer aos gerentes e analistas de negócios a capacidade de realizar análise adequada. (...) O processo de BI baseia-se na transformação de dados em informações, depois em decisões e finalmente em ações.”

As principais vantagens de uma ferramenta de BI são:

- Flexibilidade e capacidade de customização;
- Interações intuitivas com o usuário;
- Painéis analíticos interativos;
- Possibilidade de contribuição em ambientes colaborativos;
- Interface local e na nuvem;
- Facilidade de desenvolvimento para usuários comuns;
- Disseminação da informação com velocidade.

Analizando as definições explanadas acima, utiliza-se para este trabalho, um conceito de BI que engloba as ideias em comum e é pertinente com o projeto aqui desenvolvido. Esta abordagem escolhida interpreta que uma ferramenta de BI deve oferecer suporte para tomadas de decisão, sendo esta a principal característica, deve fornecer conhecimento para o controle de processo do negócio, sendo simples, claro, flexível, oferecendo velocidade e segurança, fortalecendo a cadeia de valores da organização e gerando vantagem competitiva. Estes dois últimos itens serão abordados posteriormente nesse capítulo.

3.1.2 Histórico do Business Intelligence

O surgimento do termo Business Intelligence é atribuído, em grande parte da literatura, à empresa Gartner Group, próximo da década de 90, embora diversas aplicações na área de sistemas de informação já existissem, e haviam criado a base para o BI que conhecemos hoje. Por outro lado, de acordo com Elena (2011), o termo já existia na década de 50 e foi citado pela primeira vez por Hans Peter, pesquisador da IBM (International Business Machines Coporation), no artigo “A Business Intelligence System”. Outros autores afirmam que o BI data de muito antes disso, quando os povos egípcios e persas já utilizavam informações de comportamentos da natureza, tais como posição das marés, período de chuvas e estiagens, para tomar decisões e gera valor para seus negócios.

Embora as referências do início exato do BI sejam difusas, é importante acompanhar a evolução desse segmento até o que conhecemos hoje e entender o que ainda pode se conquistar.

De acordo com Turban (2009), no início dos anos 80, com o nascimento dos sistemas de informação executivas, os negócios começaram a abranger relatórios mais dinâmicos, contendo mais análises e previsões, frente à um maior suporte de recursos de computação. Antes disso, Luhn (1958) cita também que os relatórios precisavam ser impressos e a digitalização da informação era um grande obstáculo.

O BI obteve sua possibilidade de crescimento e melhoria a partir do surgimento dos gerenciadores de banco de dados, aliado à evolução da capacidade de processamento e do avanço de mercado dos computadores pessoais (ELENA, C., 2009). Neste contexto, Barbieri (2011) também traz uma perspectiva bastante interessante. Ele explica que, na década de 80, os dados tiveram sua época de

estrelato, porém ainda somente no campo tecnológico, com pouca integração aos negócios. Desta forma, o autor atribui à internet a grande expansão das ferramentas e metodologias do Business Intelligence. A revolução dos dados proporcionada pela internet foi, diferente da Revolução Industrial, democrática, invasiva e de amplo alcance. Ela permitiu o acesso a uma aldeia global, contendo uma infinidade de informações. As empresas foram então induzidas a garantir a seus clientes e consumidores maior qualidade, opção, velocidade e menor preço. Era necessário entregar a informação mais precisa aos tomadores de decisão.

As primeiras versões de tecnologia de Business Intelligence, muitas vezes denominada BI 1.0, tinham interfaces mais complexas, com pouca facilidade de uso dos clientes finais. Era preciso conhecimento de mecanismos computacionais para manipulação das ferramentas, restringindo seu uso aos profissionais de TI. Os executivos recebiam os relatórios com informações consolidadas, de modo que a qualidade e velocidade da informação se tornavam reféns da relação entre profissionais de diferentes áreas. Os softwares eram desenvolvidos para uso local, fora da nuvem, utilizando linguagens como Java, C++ e Visual Basic. Já o BI 2.0, tem expandido seu público alvo, abrangendo profissionais sem experiência em ciência de dados ou sistemas de informação. Além disso, as interfaces têm se tornado mais simples e claras, prezando pela qualidade da informação, com migração de suas atividades para o ambiente em nuvem, o qual permite colaboração mais interativa entre equipes e facilidade no acesso. Esse movimento apresenta grande tendência de intensificação com o intenso uso das tecnologias móveis, e o avanço do movimento denominado “Indústria 4.0”.

O acompanhamento dessa nova tendência de mercado traz vantagens de maior praticidade de uso, sendo possível o acesso à informação de qualquer lugar do mundo, velocidade na tomada de decisão, abertura para ambientes mais colaborativos, integração com tecnologias de Inteligência Artificial e Aprendizado de Máquinas, etc. Entretanto, as barreiras que podem impedir o BI de se fortalecer, principalmente no ambiente industrial são o poder de processamento de servidores, o armazenamento e a segurança dos dados.

3.1.3 Business Intelligence e a Cadeia de Valor

Toda organização tem sua visão voltada para criação de valor, em busca de vantagem competitiva, o que quer dizer que estão em constante busca por crescimento, maximização da qualidade de seus produtos, melhoria contínua, redução de custos e perdas, etc. Desta forma, na era da nova economia global, líderes corporativos são cada vez mais pressionados a entregarem inovação, diferenciação e desempenho. Para isso, investem em ferramentas de BI pois entendem que a informação tem peso significativo na criação de valor de seus negócios, a partir da possibilidade do monitoramento e gestão da performance, permitindo tomadas de decisão mais efetivas.

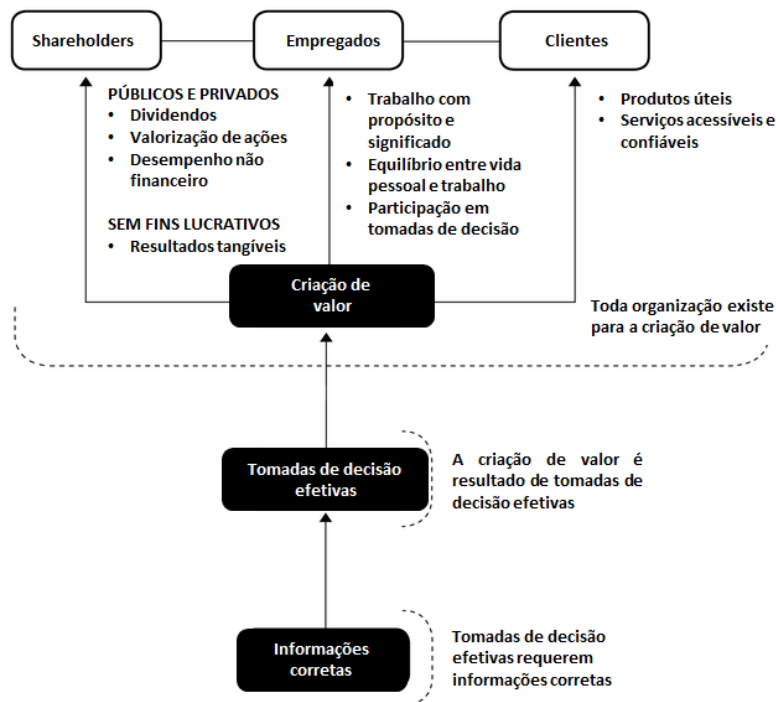
Para entender essa relação, Steffine (2015) expõe que a criação de valor nas organizações está diretamente ligada à compreensão dos interesses de seus stakeholders, ou seja, dos indivíduos e grupos que podem afetar ou são afetados pelo negócio (tais como investidores, funcionários e clientes).

No entanto, essa criação de valor é resultado direto de tomadas de decisão inteligentes e baseadas em fatos. Organizações que tomam decisões rápidas e assertivas estão à frente de mercado e obtêm vantagens, mesmo sobre grandes corporações. O autor também destaca a mudança de cenário corporativo no qual os “grandes” dominavam sobre os “pequenos” para um universo onde os “mais rápidos” dominam aqueles que não o são.

Desta forma, a efetividade das tomadas de decisão requer informações corretas e relevantes. Isso quer dizer que, nem todo dado é importante no contexto do BI. O grande volume de dados disponíveis deve ser cuidadosamente filtrado, tratado e analisado, para garantir que a informação gere conhecimento e, consequentemente, ações assertivas.

A Figura 1 resume a cadeia de valores resultante do Business Intelligence.

Figura 1: Cadeia de Valores do BI



Fonte: STEFFINE, 2015.

3.1.4 Ciclo da Informação e Arquitetura do Business Intelligence

Para compreender o ciclo da informação e como ele funciona em um ambiente de Business Intelligence, é necessário antes distinguir dados, informação e conhecimento. Os dados são a parte mais bruta, são registros de acontecimentos, especificações de produtos, etc. Sem o devido tratamento, dados não representam muito para uma organização. A informação, no entanto, é a parte inteligente, consolidada e precisa dos dados. O conhecimento são os insights para tomada de decisões, proporcionados pela informação compilada dos dados. Um exemplo dessa diferenciação para um lojista seriam dados como: “Produto A, Marca M1, Parâmetros P1; Produto A, Marca M2, Parâmetros P1; Produto A, Marca M2, Parâmetro P1”. Como informação poderia ser concluído: “1500 unidades de Produto A vendidas em um mês, sendo 80% da quantidade total referente à Marca M1 e 75% do faturamento total concentrado na Marca M2. Predição de aumento de 20% nas vendas de M2 para o próximo mês. ” O conhecimento extraído seria: “No mês de janeiro é necessário aumentar a produção de Produto A Marca M1”. Este simples exemplo ilustra as

diferenças entre esses conceitos, enquanto que a Tabela 1 compara suas características.

Tabela 1: Comparação entre dado, informação e conhecimento

Dado	Informação	Conhecimento
Detalhe	Consolidado	Consolidado
Não apoia a decisão	Apoia a decisão	Provê insights e ideias
Operação	Gestão	Gestão
É criado a partir de uma transação	É criada a partir de um conjunto de dados	É criado a partir de um conjunto de informações

Fonte: BONEL, 2017.

Estando claro esses conceitos, é possível definir o ciclo da informação, ilustrado na Figura 2.

Figura 2: Ciclo da Informação



Fonte: BONEL, 2017.

- **Aquisição de dados:** essa é a etapa de coleta, na qual são selecionados os dados que serão insumo para o processamento da informação. Esses dados originam-se no processo operacional da organização e deverão ser armazenados em um sistema.

- Geração da Informação: esta etapa é composta por agrupamento dos dados, excluindo aqueles que não são relevantes para o negócio, ou para a área de negócio foco da aplicação
- Prover conhecimento: construção de indicadores e métricas capazes de fornecer análises para os tomadores de decisão. As informações devem estar acessíveis e dinâmicas para a formação de conhecimento.
- Decisão: de posse do conhecimento gerado pela informação, é hora de tomar decisões relevantes para o negócio.
- Ação: resultantes das decisões tomadas, para que iniciativas tenham efeito na organização.
- Acompanhar resultados: momento de mensurar o desempenho da ferramenta e sua aplicabilidade. É preciso que os gestores estejam utilizando o BI como suporte para alavancar os resultados da empresa. Portanto, nessa etapa é preciso rever os processos, encontrar possíveis erros que possam estar reduzindo a qualidade da informação

O ciclo da informação é importante para compreender a arquitetura do Business Intelligence, a qual é composta, segundo Turban (2009), por: um *Data Warehouse* (DW) ou *Data Mart* (DM), a análise de negócios, uma coleção de ferramentas de manipulação de dados, incluindo um data mining, um sistema de Business Performance Management (BPM) e uma interface de usuário. Cada um desses itens é descrito a seguir.

3.1.4.1 Data Warehouse

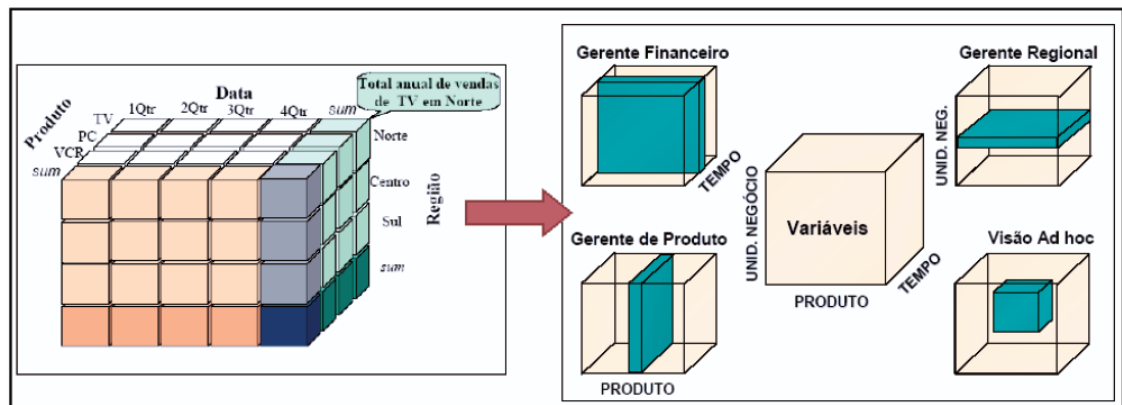
É um banco de dados suporte para tomada de decisão em formato mais simples, com a possibilidade de manipulação dos fatos. Um Data Warehouse é, portanto, um conjunto de dados orientados por assuntos, não volátil, variável com o tempo e integrado. Essas características fundamentais são descritas a seguir.

- Orientado por assuntos: os dados são organizados por subcategorias pertinentes à tomada de decisão, proporcionando uma visão mais abrangente. Exemplo: vendas, produtos, clientes;

- Integrado: o DW dá suporte para diferentes fontes de dados e supera discrepâncias de unidades de medida e conflitos de nomenclatura.
- Variável no tempo: manter uma série temporal é importante para analisar os dados ao longo do tempo, detectando tendência, variações, relações de longo prazo, previsões e comparações.
- Não volátil: dados não são excluídos ou atualizados no Data Warehouse, exceto se forem considerados incorretos ou obsoletos.

Um Data Warehouse é diferente de um banco de dados operacional, pois armazena dados gerenciais, que contêm em si informação útil capaz de gerar conhecimento. Para permitir consultas rápidas e eficazes durante o processo de análise, o DW agrupa dados frequentes em formato de cubo multidimensional. Esse modelo deriva da ideia que as informações se relacionam de maneira que podem ser representadas como um cubo. O cubo pode ser fatiado, de modo que cada eixo permite extrair visões diferentes sobre a informação. A figura a seguir ilustra um exemplo dessa característica multidimensional.

Figura 3: Cubo da Informação



Fonte: ANTONELLI, 2009.

Outra característica importante é que os DW's contêm metadados, ou seja, dados sobre dados, a fim de organizá-los e utilizá-los de maneira mais eficiente.

3.1.4.2 Data Mart

Os Data Marts possuem as mesmas características de um DW, diferenciando apenas o fato que se concentram em um assunto ou departamento específico, um subconjunto. Ele pode ser dependente, ou seja, criado a partir do Data Warehouse e utilizando um modelo consistente e já testado pela organização, ou independente, projetado para uma pequena unidade estratégica de negócios ou um departamento, mas que não deriva de um DW.

Para consolidar a diferença entre DW's e DM's a Tabela 2 abaixo faz um comparativo entre suas modelagens.

Tabela 2: Comparação entre Data Marts e Data Warehouse

Data Marts	Data Warehouse
Nível departamental	Nível corporativo
Alto nível de granularidade	Baixo nível de granularidade
Pequena quantidade de dados históricos	Grande quantidade de dados históricos
Tecnologia otimizada para acesso de consultas rápidas	Tecnologia otimizada para armazenamento e gerência de grandes quantidades de dados
Cada área departamental possui suas características específicas	As estruturas são reconstruídas para um entendimento em nível de corporação

Fonte: RIBEIRO, 2005.

3.1.4.3 OLAP x OLTP

Dois conceitos muito utilizados em BI são o “Online Analytical Processing” (OLAP) e o “Online Transaction Processing” (OLTP). Embora os dois termos estejam ligados à área da tecnologia de dados, seus propósitos são distintos. O OLTP é voltado para operações, com bases de dados dinâmicas, sendo capaz de gerar relatórios atualizados, mas nunca uma base histórica. Ele está voltado para a organização e categorização de informações, de modo que possui alta velocidade de manipulação de dados. Já o OLAP está voltado para análises, insights, permitindo que o usuário interaja com diferentes pesquisas. Seu foco é em nível estratégico, tem leitura otimizada, com alto nível de sumarização. É armazenado em DW's e sua atualização é feita no momento de carga dos dados, sendo uma base histórica, não volátil. A Tabela 3 a seguir ilustra as diferenças entre os dois sistemas.

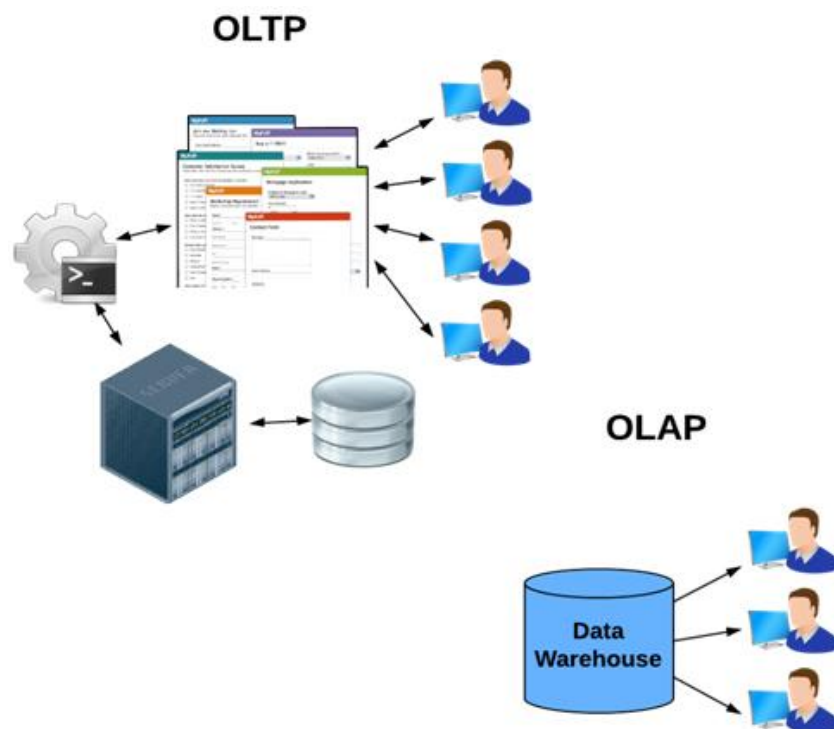
Tabela 3: Comparação entre OLTP e OLAP.

OLTP	OLAP
Voltado para operações dia a dia	Voltado para performance analítica
Baixa performance em consultas	Alta Performance em consultas
Modelagem ramificada	Modelagem simplificada (star)
Histórico de operações inexistente	Armazém de dados (Histórico existente)
Volátil	Não volátil

Fonte: RIBEIRO, 2005.

Analisando os dois conceitos, não se pode dizer que um seja melhor que o outro. Porém, no universo do BI, as aplicações utilizam o modelo OLAP. A próxima figura ilustra que os usuários conseguem realizar transações que alteram o banco de dados no modelo OLTP, enquanto podem apenas realizar consultas no modelo OLAP, razão principal da exclusividade do OLAP em sistemas para tomada de decisão.

Figura 4: Modelos OLTP e OLAP.



Fonte: ELIAS, 2019.

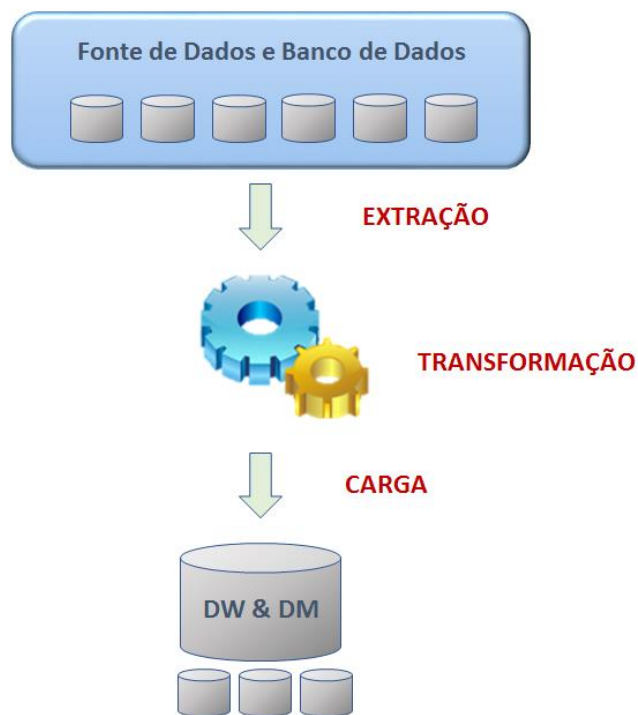
3.1.4.4 ETL (Extração, Transformação e Carga)

De acordo com Barbieri, o processo de ETL baseia-se em 5 passos:

- Identificação da origem dos dados;
- Limpezas, ajustes, correções de imperfeições;
- Padronização e correção de variabilidades;
- Carregar dados para o DW;
- Atualização do DW.

Processos de ETL são fundamentais em ambientes de BI, pois proporcionam dados de qualidade para os DW's e DM's. Desta forma, ele deve ser manutenível e escalável, admitindo a escolha do momento de execução de suas etapas. Ele permite a consolidação de dados oriundos de fontes externas diferentes, unificando e aplicando regras comuns que agreguem valor e gerem conhecimento para a organização. A Figura 5 abaixo apresenta o esquema.

Figura 5: Processo de ETL.



Fonte: Autor, 2019.

3.1.4.5 Data Mining

Data Mining significa exploração de dados e deriva da concepção de identificação de padrões nos dados operacionais, ou derivados de um DW ou DM. De acordo com Turban (2009):

“Data Mining é um processo que usa técnicas estatísticas, matemáticas, de inteligência artificial e de aprendizagem automática para extrair e identificar informações úteis e conhecimento subsequente de dados”.

Estão incluídas as tarefas de extração de conhecimento, arqueologia e exploração de dados, processamento de padrões, limpeza de dados irrelevantes, etc.

Os principais modelos de Data Mining são:

- Modelos simples: consultas baseadas em SQL, processamento analítico online (OLAP);
- Modelos intermediários: regressão, árvores de decisão, agrupamento;
- Modelos complexos: redes neurais

De maneira geral, classifica-se o Data Mining em três grandes vertentes:

- Estatística: é o pilar mais antigo e consolidado do Data Mining, sendo a base para a maioria das análises a nível de usuário.
- Inteligência Artificial: construída a partir de fundamentos heurísticos, fundamenta atuar de maneira semelhante ao pensamento racional humano na resolução de problemas estatísticos.
- Machine Learning: é composto pela combinação da estatística com a fundamentação heurística, no qual os computadores aprendem com base em dados, automatizando o reconhecimento de padrões complexos.

Observa-se, portanto, a grande importância do Data Mining, sobretudo na era da Inteligência Artificial e Aprendizado de Máquinas, conquistando espaço de mercado e tornando-se peça sucesso e sobrevivência das organizações na era da informação.

3.1.5 Aplicações de BI nas organizações

As organizações utilizam o BI com objetivos específicos, alguns descritos a seguir:

- **Mensuração:** recebendo entradas de dados referentes à itens de controle do processo empresarial, as aplicações podem trabalhar com indicadores por meio destes dados;
- **Análise:** por meio dos dados, análises são geradas, buscando tendências e conhecimento significativo;
- **Relatórios:** com o uso de BI, a geração de relatórios personalizados é um processo muito mais rápido e fácil, substituindo planilhas e programas de texto.
- **Colaboração:** relatórios podem ser criados e editados em nuvem, em um ambiente colaborativo, permitindo participação de diversos colaboradores, sem exigir que estejam trabalhando no mesmo ambiente.

As aplicações de BI na atualidade atingem os três níveis hierárquicos: estratégico, tático e operacional, sendo este último uma novidade no mundo dos negócios. A atuação nesses níveis deve ser integrada, de modo que o BI estratégico dá dinâmica ao BI tático e direciona o operacional. Este primeiro nível é focado no monitoramento do desempenho e tem como objetivo impulsionar a performance geral da organização. Nesse segmento trabalha-se com indicadores e mapas estratégicos, voltados para longo prazo. Já o BI analítico busca a identificação e resolução de problemas que podem impactar os indicadores de desempenho, dirigindo as iniciativas operacionais. Este é, portanto, focado em domínios da organização, tais como áreas de marketing, vendas, etc. Por fim, o operacional dá autonomia aos funcionários a buscarem soluções imediatas, automatizando processos e atuando no curto prazo. A Tabela 4 reúne as principais características descritas dos três níveis.

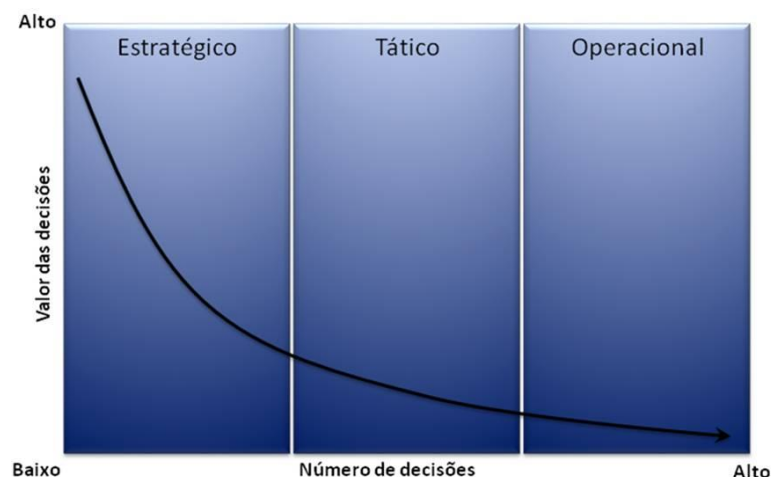
Tabela 4: Características do BI em seus diferentes níveis.

Característica	BI Estratégica	BI Tática	BI Operacional
Foco principal do negócio	Attingir as metas empresariais em longo prazo	Analisar dados; entregar relatórios	Administrar operações do dia a dia com relação a atingir metas
Principais usuários	Executivos, analistas	Executivos, analistas, gerentes de setor	Gerente de setor
Métricas	Métricas são um mecanismo de <i>feedback</i> para acompanhar e entender como a estratégia está progredindo e quais ajustes precisam ser planejados	Métricas são um mecanismo de <i>feedback</i> para acompanhar e entender como a estratégia está progredindo e quais ajustes precisam ser planejados	Métricas são individualizadas para que o gestor de cada linha possa obter <i>insight</i> sobre o desempenho de seus processos de negócio
Prazo	Mensal, trimestral, anual	Diário, semanal, mensal	Imediatamente, dentro do dia
Tipos de dados ou usos	Histórico, preditivo	Histórico, preditivo	Em tempo real ou quase em tempo real

Fonte: BOTELHO E FILHO, 2014.

Estes três níveis carregam consigo diferentes formas de tomada de decisões. No nível estratégico, as decisões são de alto valor e realizadas de maneira individual, com baixa frequência. Decisões de nível tático são de médio valor, possuindo maior volume e frequência, enquanto as operacionais são de alto volume em uma pequena escala de tempo, apresentando baixo impacto quando analisadas separadamente, porém diretamente ligadas ao objetivo central da organização quando observadas em conjunto. A Figura 6 mostra a curva que ilustra a relação entre o número de decisões e o valor destas para o negócio.

Figura 6: Relação de número de decisões versus o valor das decisões para o negócio.



Fonte: GUEDES, 2013.

O projeto proposto neste trabalho tem seu direcionamento muito voltado para o nível operacional e tático, sendo o operacional de maior desafio, uma vez que exige

uma estrutura de automação, coleta e armazenamento de dados capaz de entregar a informação muito próxima do tempo real.

3.1.6 Self-Service BI

Self-Service BI é uma forma de autosserviço que dispensa profissionais especializados em sistemas de informação, permitindo que usuários comuns sejam capazes de desenvolver toda a estrutura de uma aplicação. Ferramentas de Self-Service BI são totalmente intuitivas e desenvolvidas para que qualquer usuário seja capaz de criar análises personalizadas em pouco tempo. Sua construção permite que não seja necessária toda a arquitetura de BI descrita anteriormente e, assim, usuários comuns de pequenas organizações, ou mesmo empreendedores individuais sejam capazes de disfrutar dos benefícios da ferramenta. Planilhas de Excel e arquivos de texto podem ser conectadas como fonte de dados.

Além disso, o custo de licenciamento de soluções em Self-Service é muito menor, os aplicativos podem ser baixados e testados em versão teste e libera os profissionais de TI para um trabalho mais efetivo e especializado para a organização. Desta maneira, a escolha de uma solução como esta pode ser de bastante sucesso para o ambiente de negócios.

3.1.7 Soluções de BI de mercado

Existem hoje no mercado diversas soluções de tecnologia de Business Intelligence disponíveis, as quais oferecem aos profissionais diversas possibilidades para projetos integrados no setor. Para uma boa escolha na hora de adquirir uma destas soluções é importante levar em consideração a velocidade de processamento de dados, estabilidade de performance, facilidade de uso e custo-benefício.

A Figura 7 é conhecida como Quadrante Mágico e ilustra o posicionamento das soluções de mercado mais relevantes. Esse quadrante é proposto pela Gartner Group e representa uma avaliação das capacidades destes produtos. Seu eixo horizontal representa a amplitude da visão da empresa em relação à tecnologia, enquanto que o eixo vertical a sua capacidade de entrega do que se propõe. A Gartner define que os líderes possuem alta visão de inovação, criando valor para seus clientes e possibilitando fácil aplicação.

Figura 7: Quadrante Mágico das soluções de BI de mercado.



Source: Gartner (February 2019)

Fonte: Gartner Group, 2019.

Os dois maiores líderes destes produtos são apresentados a seguir, para conhecimento das principais ofertas de soluções de BI atuais.

3.1.7.1 Microsoft Power BI

A Microsoft é líder no ranking de avaliação da Gartner por 12 vezes consecutivas, sucesso que se deve ao forte investimento que a plataforma aplica ao produto. O produto foi lançado em 2015 e conta com diversas funcionalidades tais como serviço para Desktop, nuvem e suporte para smartphones. A ferramenta pode

ser utilizada na sua versão gratuita, Power BI Desktop, ou nas suas funcionalidades pagas, Power BI Pro e Power BI Premium. A versão Pro é de custo reduzido e traz a proposta de um ótimo custo-benefício para implantar o BI no ambiente organizacional. A maior proposta é trazer inovação rápida e orientada ao usuário, além de atualizações constantes de software que melhoras os recursos e as funcionalidades do programa. Essas atualizações são resultados de insights e inputs da própria comunidade usuária da ferramenta, por meio de seu blog de aprendizado. É importante lembrar que o Power BI é uma proposta de self-service BI, ou seja, o autosserviço, no qual o usuário comum é capaz de criar seu desenvolvimento em BI sem a necessidade de profissionais especializados na área.

Figura 8: Microsoft Power BI



Fonte: Microsoft, 2019.

Em suma podemos destacar como vantagens do Power BI:

- Possibilidade de conexão à fontes de dados variadas;
- Fácil integração com diferentes dispositivos tais como computadores pessoais, tablets e smartphones;
- Serviço em nuvem que permite o desenvolvimento colaborativo;
- Custo reduzido para versão Pro;
- Visuais customizáveis, tratamento de dados simplificado;

- Suporte para mapas tridimensionais.

Suas desvantagens são:

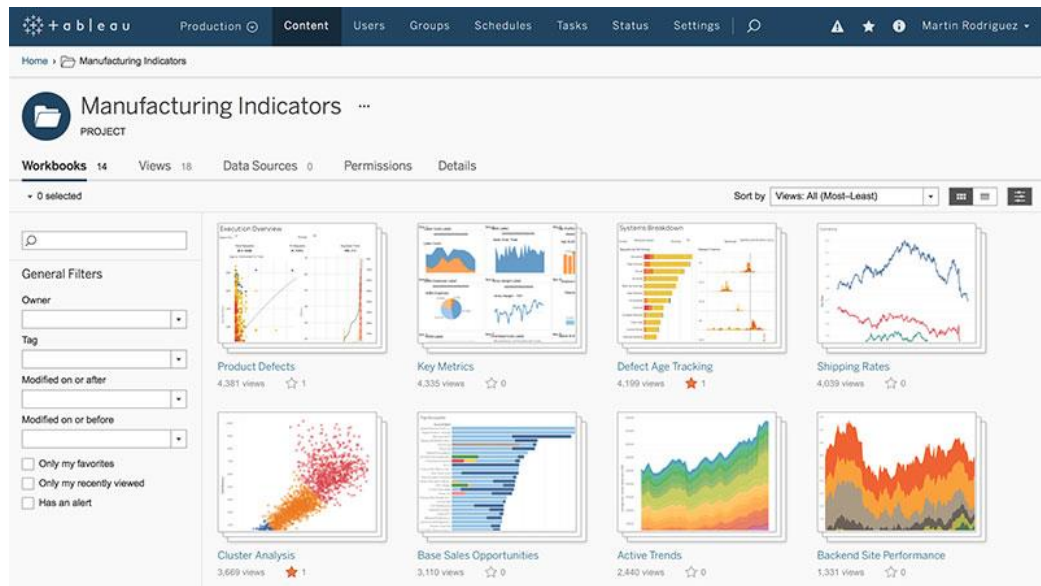
- Dependência de um bom recurso de processamento de dados;
- Desempenho para grandes volumes de dados ainda incipiente;
- Funcionamento modular demanda tempo para entendimento de cada parte do todo;
- Erros em consultas podem causar problemas em todo o restante do ambiente;
- Agendamento de reports, atualização automática e notificações personalizadas ainda pouco desenvolvidas.

Tendo em vista os pros e contras apresentados, o Power BI foi escolhido para o desenvolvimento da ferramenta de BI apresentada neste trabalho. Suas principais funcionalidades são explicadas posteriormente, integrado aos objetivos centrais do projeto.

3.1.7.2 Tableau

Logo após o Power BI, o Tableau é uma das ferramentas apontadas como mais poderosas no mercado de BI. É uma das mais utilizadas no mundo, famosa por sua ótima capacidade de visualização de dados. Foi lançada em 2003, anterior à sua concorrente Microsoft, o que lhe possibilitou vantagem competitiva de liderança por uma década antes do lançamento do Power BI. Seu maior diferencial quando comparado ao Power BI é sua robustez. O Tableau não é uma ferramenta para somente visualização de relatórios, oferecendo suporte maior para grandes volumes de dados. Sua implantação é simplificada, porém não tanto quanto o Power BI, que oferece a possibilidade de criar um dashboard interativo em minutos. Com relação à análise de dados, o Tableau possui vantagem em fornecer recursos extensivos, permitindo análise de tendências, muito voltado para profissionais de Data Science. Além disso, o custo de licença do Tableau pode chegar em até dez vezes mais que o Power BI, impeditivo para pequenas e médias empresas iniciantes na área de inteligência de dados.

Figura 9: Tableau



Fonte: Tableau, 2019.

Para concluir a comparação entre essas duas soluções a Tabela 5 compila as vantagens de cada um deles:

Tabela 5: Power BI x Tableau

Item	Power BI	Tableau
Ano de lançamento	2013	2003
Aplicação	Dashboards e Visualizações	Análise de Dados
Custo	Baixo	Alto
Usuários	Técnicos / Não-técnicos	Analistas
Nível de Suporte	Baixo	Alto
Escalabilidade (com grandes conjuntos de dados)	Boa	Muito boa
Licenciamento	Rígido	Flexível
Infraestrutura	SaaS	Flexível
Funcionalidades em Geral	Bom	Muito bom

Fonte: MATOS, 2019.

3.1.8 Futuro do BI

Toda a fundamentação teórica exposta neste capítulo permite concluir que o Business Intelligence tem interface cada vez maior com tecnologias e inovações que vêm crescendo e ganhando força no mundo corporativo. Em um cenário em que é preciso aumentar a velocidade e a assertividade das tomadas de decisão, utilizar a tecnologia dos dados é chave para o sucesso organizacional, ampliando vantagem

competitiva de mercado. O BI tem acompanhado essa tendência e integrado funções inovadoras às suas aplicações. Os programas mais modernos de BI integram o Machine Learning para reduzir o volume de dados que não agregam decisões assertivas, uma vez que, à medida que os dados são inseridos, os algoritmos se adaptam e otimizam as análises, liberando tempo para os usuários pensarem mais estrategicamente e menos analiticamente.

Outro exemplo disso é a utilização de Processamento de Linguagem Natural (PNL) na extração de dados não estruturados da linguagem escrita ou falada, computando-os e apresentando respostas aos usuários antes mesmo que as perguntas sejam feitas. Além disso, o aumento da força da IoT (Internet of Things), uma tendência da tecnologia dos dados integrada com as aplicações em nuvem, marca um avanço enorme na produção, coleta e visualização dos dados. Outro fator importante será o aumento do uso do Self-Service BI, retirando a necessidade de profissionais qualificados tecnicamente em ciência de dados ou sistemas de informação.

Pesquisas mostram que as organizações deverão utilizar o Business Intelligence se desejarem aumentar seus lucros e receitas. Essa realidade é confirmada pelo crescente aumento dos investimentos ao redor do mundo. A previsão para 2020 é que o BI alcance os U\$26,7 bilhões, representando um crescimento anual de 8,4%, com a entrada em massa de bancos, seguradoras, varejistas, empresas de TI e telecomunicações.

Toda a expectativa sobre o BI para o futuro acarretará em mudanças em toda a organização, seja no fluxo de tomada de decisões, seja na capacitação de pessoas, ou na forma como os processos são conduzidos.

3.2 Machine Learning

3.2.1 Conceitos

O termo Machine Learning é outro muito utilizado para abordagens diferentes, em contextos diversos, principalmente no momento atual em que a disseminação da aposta nesta tecnologia está em alta. Uma definição bastante interessante, apontada como primeira aparição do termo, diz tratar-se de um campo de estudo que possibilita que computadores aprendam sem serem explicitamente programados (Samuel, A., 1959).

Da definição em português, Aprendizado de Máquinas, é considerado em vários estudos como uma subárea da Inteligência Artificial que, conforme explica Goodfellow et al.(2016), procura prover conhecimento aos computadores por meio de dados, observações e interações com o mundo.

Entretanto, para compreender bem o que é esta ciência, o conceito mais direcionado seria, de acordo com Tom Mitchell (1997):

“Um programa aprende com a experiência E em relação a uma determinada tarefa T e uma medida de desempenho P, se o seu desempenho na tarefa T, medido por P, melhora com a experiência E. ”

Desta maneira, o conhecimento adquirido pelas máquinas deve ser capaz de ser generalizado para novas situações, ou seja, novos dados, ainda não conhecidos. Para tal, os dados utilizados no treinamento da máquina devem ser suficientemente representativos, não contendo ruídos e valores incompatíveis com a situação de aprendizado.

Em concordância a estes conceitos, o principal objetivo das técnicas que Machine Learning são o entendimento e aplicação de métodos utilizando algoritmos capazes de realizar tarefas automaticamente sem assistência humana. Seu maior foco é o desenvolvimento de programas capazes de se adaptarem à exposição de novos dados.

Desta maneira, as principais vantagens do Aprendizado de Máquinas são:

- Capacidade de aprendizado e melhoria com seus próprios erros;
- Velocidade de análise e resultados;
- Melhoria na gestão de dados;
- Permite automatização de processos;

- Apresenta soluções para problemas reais, tais como análise de riscos em ambiente de trabalho;
- Reduz custos, uma vez que fornece informações reduzidas de erros, evitando desperdícios, melhorando desempenhos, etc;

Diversas pesquisas endossam tais vantagens ao apontar o Aprendizado de Máquinas como grande tendência de inclusão nas empresas nos próximos anos, em conjunto com as áreas de Business Intelligence, Big Data e Internet das Coisas. Nos últimos anos têm se observado uma crescente utilização das técnicas de Aprendizado de Máquinas para a análise do comportamento de consumidores, aplicando sugestões de produtos para compra e entregando estratégias de marketing e vendas muito mais eficazes.

Apesar das inúmeras aplicações e vantagens descritas, Machine Learning ainda enfrenta algumas dificuldades para sua consolidação efetiva, tais como:

- Disseminação de informações distorcidas e enviesadas sobre as técnicas de Machine Learning, denegrindo o real potencial dessa área;
- Grande volume de dados disponível com baixa qualidade, apresentando ruídos e com grande necessidade de tratamento;
- Carência de profissionais capacitados na área;
- Utilização de dados não relacionados à área de estudo e aplicação do Aprendizado de Máquinas;
- Falta de planejamento na implantação de projetos;
- Rejeição à novas tecnologias por parte das equipes de trabalho nas empresas.

Baseado nos conceitos apresentados e no projeto desenvolvido, para este trabalho, Aprendizado de Máquinas é abordado como um conjunto de técnicas de criação de algoritmos capazes de aprender padrões e criar modelos a partir de dados, e realizar descrições e previsões sobre valores desconhecidos, gerando aplicabilidade para solução de problemas.

3.2.2 Histórico do Aprendizado de Máquinas

Embora o termo Machine Learning esteja em seu auge, o conceito de Aprendizado de Máquinas associado com a capacidade de detectar automaticamente

padrões nos dados remete sua existência desde os anos 50, período em que diversos métodos estatísticos e matemáticos foram aprimorados no meio acadêmico. Foi nesta década que Arthur Samuel, cientista da IBM (International Business Machine), lançou o primeiro programa de aprendizado para jogo de damas, o qual melhorava seus algoritmos à medida que jogava, estudando movimentos e propondo estratégias vencedoras, incorporando os aprendizados ao seu escopo. Neste mesmo período, surgia também o Perceptron, criado por Frank Rosenblatt. Trata-se de um tipo de rede neural que introduziu os conceitos em termos matemáticos e permitiu o avanço dessas técnicas para o que conhecemos hoje.

Já nos anos 60 surgiram os primeiros modelos de reconhecimento de padrões, com técnicas como a de utilização de métricas de distância Euclidiana, muito utilizadas até hoje em Aprendizado de Máquinas.

O início do despontamento do Aprendizado de Máquinas aconteceu de fato na década de 90, quando os algoritmos passaram a ser aplicados em conjunto com técnicas de Data Mining, em aplicações Web, desenvolvendo avanços numerosos para essa ciência. A partir daí, novas formas de avaliar o desempenho dessas técnicas também passaram a ser utilizadas, estimando a precisão dos algoritmos.

Em suma, é possível compreender que, apesar de seus conceitos não serem recentes, o Aprendizado de Máquinas alcançou crescimento significativo apenas nas últimas décadas, com a possibilidade oferecida pelas novas tecnologias, pela internet e pela melhoria contínua na capacidade de processamento das máquinas. O que se pode apontar é que essa área apresentará ainda mais êxito, com o aperfeiçoamento de algoritmos e surgimento de profissionais ainda mais habilitados.

3.2.3 Tipos de Aprendizado de Máquinas

Conforme será detalhado posteriormente, existem diversos tipos de algoritmos de Machine Learning, de modo que uma classificação usual é baseada nos seguintes critérios:

- Existência de uma supervisão humana no treinamento ou não:
Aprendizado Supervisionado, Não Supervisionado, Semisupervisionado ou Aprendizado por Reforço;

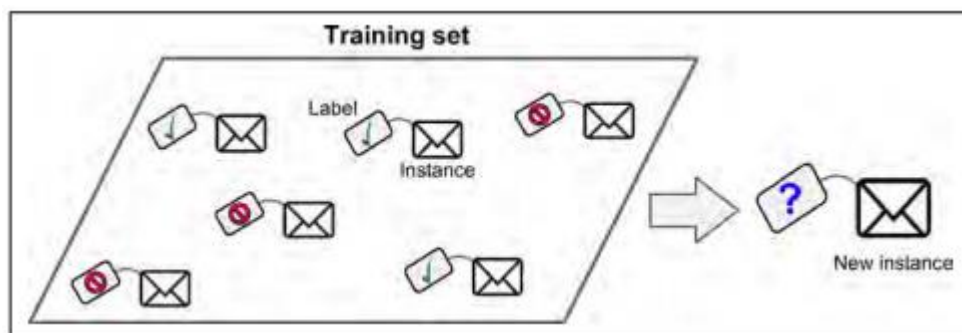
- Possibilidade de aprendizado incremental online ou não: **Aprendizado online ou por batelada;**
- Comparação de novos dados com dados conhecidos, ou detecção de padrões nos dados de treinamento, criando um modelo preditivo: **Aprendizado baseado em instância ou modelo.**

Os parâmetros apresentados não são exclusivos, isto é, um mesmo algoritmo pode ser classificado com tais critérios combinados. Os detalhes de cada um desses tipos são apresentados a seguir.

3.2.3.1 Aprendizado Supervisionado e Não Supervisionado

No aprendizado supervisionado, o conjunto de dados de treinamento inclui a solução desejada, chamada de “label”, ou em português, “rótulo”. Esse caso é muito aplicado em classificações e regressões. A Figura 10 ilustra esse conceito.

Figura 10: Exemplo de Aprendizado Supervisionado.



Fonte: GÉRON, 2017.

Exemplo disso são algoritmos para filtros de spam em e-mails: um agrupamento de dados referentes à vários e-mails classificados ou não como spam é fornecido ao treinamento, e espera-se classificar novos e-mails. Outro exemplo seria a predição do valor de carros, de acordo com características, como ano de fabricação, quilometragem, marca, etc, em que o algoritmo de treinamento recebe um conjunto de dados contendo tais características e seus rótulos (preços).

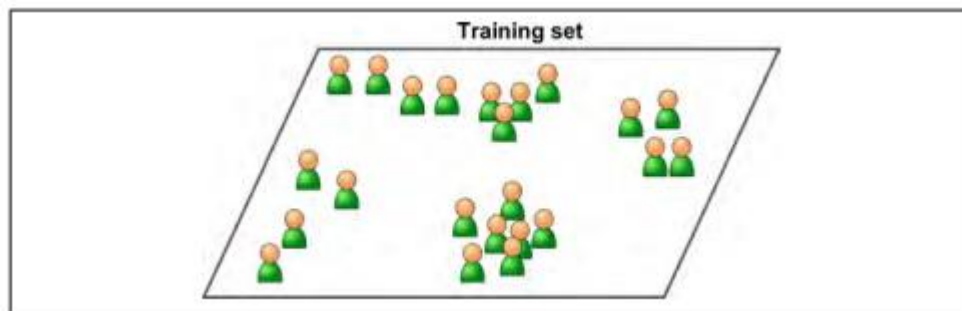
Os algoritmos de Aprendizado Supervisionado mais importantes são:

- Regressão Linear;

- Regressão Logística;
- K-vizinhos mais Próximos;
- Máquinas de suporte Vetorial (SVM's – Support Vector Machines);
- Árvores de Decisão;
- Redes Neurais Artificiais.

Por outro lado, com relação ao Aprendizado Não Supervisionado, não há uma inclusão do rótulo no conjunto de dados de treinamento, conforme mostra a Figura 11.

Figura 11: Exemplo de Aprendizado Não Supervisionado.



Fonte: GÉRON, 2017.

Deste modo, no Aprendizado Não Supervisionado o computador precisará identificar perfis nos dados, encontrando uma representação mais informativa, simples e condensada. Um exemplo muito utilizado nos dias atuais é a recomendação de séries, filmes e músicas nos aplicativos e plataformas de entretenimento, baseada em perfis de utilização do usuário.

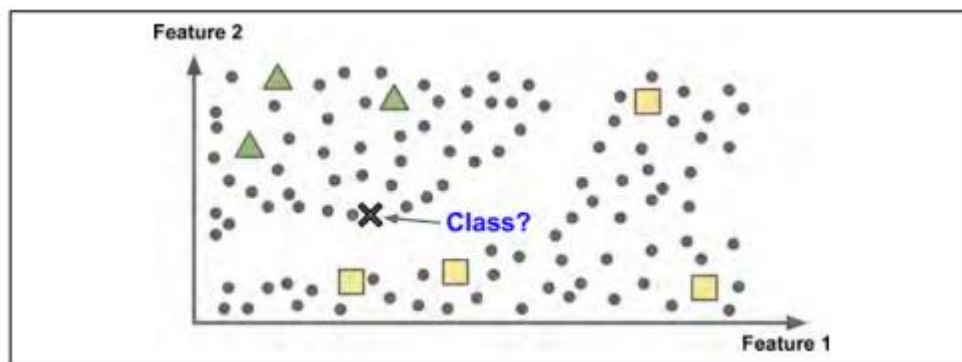
Alguns dos algoritmos de Aprendizado Não Supervisionado:

- K-Nearst Neighbors (K-Médias);
- Clusterização Hierárquica;
- Expectativa-Maximização;
- Máquinas de Bolstzman
- T-SNE (t-distributed Stochastic Neighbor Embedding).

Paralelamente a esses conceitos, alguns algoritmos podem apresentar os dois comportamentos, contendo um grande volume de dados não supervisionados e

alguns contendo seus respectivos rótulos supervisionados. É o caso do aplicativo Google Photos, no qual o usuário faz o upload de várias fotos, o sistema é capaz de automaticamente reconhecer que uma mesma pessoa aparece em fotos diferentes, com outras pessoas, realizando uma clusterização destas informações. Posteriormente, o usuário alimenta o sistema dizendo em apenas uma foto quem é aquela pessoa, dando a capacidade de o algoritmo classificar todas as outras. Este esquema é representado na Figura 12.

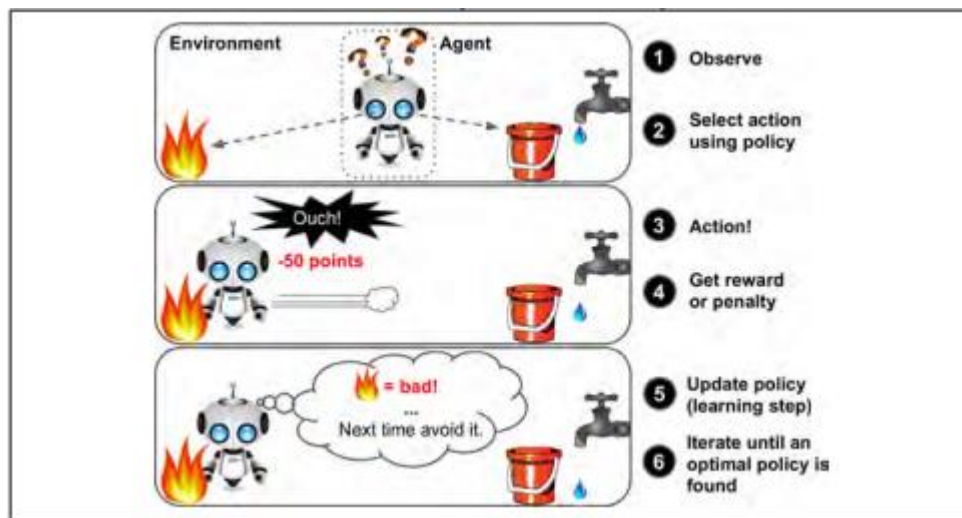
Figura 12: Exemplo de Aprendizado Semisupervisionado.



Fonte: GÉRON, 2017.

Por fim, o Aprendizado por Reforço pode ser caracterizado como um sistema de recompensas, em que a máquina observa as circunstâncias, toma diversas ações, recebendo uma pontuação para cada uma delas. Esse aprendizado é derivado da aprendizagem por reforço estudada pela área de psicologia, em que uma recompensa ou penalidade é dada a um agente, dependendo da decisão tomada. Ao longo do treinamento, espera-se que o agente saiba qual a ação que gera mais recompensas e menos penalidades. Esse esquema é mostrado na Figura 13.

Figura 13: Exemplo de Aprendizado por Reforço.



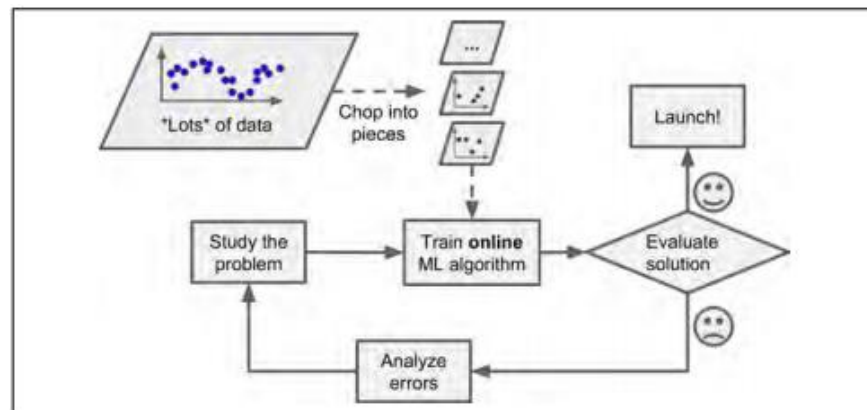
Fonte: GÉRON, 2017.

3.2.3.2 Aprendizado Online ou por Batelada

A capacidade de o sistema aprender incrementalmente, conforme novos dados são inseridos no conjunto de treinamento, ou por lotes também é importante. No aprendizado por lotes, ou batelada, a quantidade de dados é grande e, com isso, é necessário que o treinamento seja feito off-line, uma vez que necessitará de maior capacidade de processamento e tempo. Desta forma, o processo de treinamento, validação e carregamento do algoritmo pode ser feita de forma automática, substituindo o modelo anterior. As restrições para este tipo são, portanto, em sistemas que apresentam mudanças frequentes, com fluxo de dados constante, e para aplicações com fortes restrições de recursos de armazenamento e processamento.

Entretanto, existe um tipo de aprendizado, chamado de Online, em que o sistema é treinado conforme os dados são adicionados, individualmente ou por pequenos grupos. Esse tipo é ideal para sistemas que recebem dados continuamente, e precisam se adaptar a mudanças rapidamente. Além disso, é ideal para grandes volumes de dados, com poucos recursos computacionais: o algoritmo carrega parte dos dados, executa um passo de treinamento e repete o processo até que todos os dados sejam contemplados. Esse fluxo é ilustrado na Figura 14, a seguir.

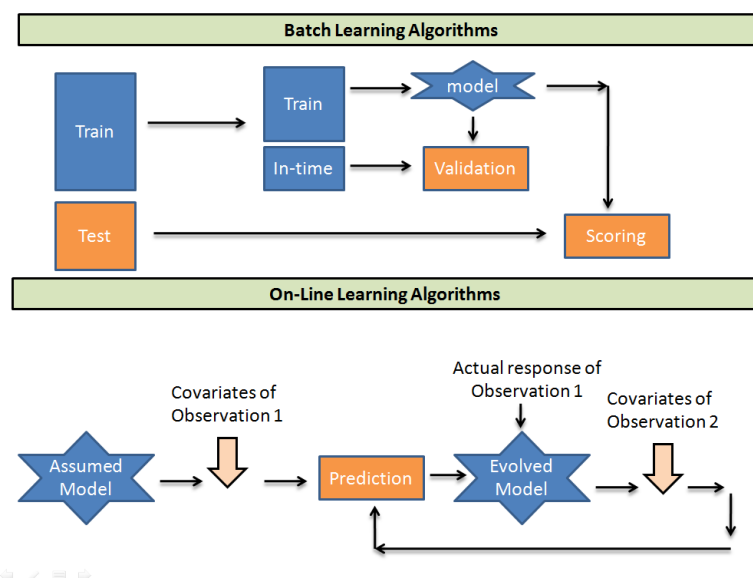
Figura 14: Aprendizado Online.



Fonte: GÉRON, 2017.

Um tópico importante para o Aprendizado Online é a taxa de aprendizado, ou seja, o quão rápido o sistema deve se adaptar às mudanças do conjunto de dados. Se a taxa de aprendizado for muito alta, o sistema irá se adaptar rapidamente, porém terá uma tendência a esquecer-se dos dados mais antigos. Por outro lado, baixa taxa de aprendizado significará em uma adaptação lenta, porém com menor sensibilidade à ruídos e dados não representativos. Uma comparação entre esses dois tipos é representada na Figura 15.

Figura 15: Comparativo entre Aprendizado Online e por Batelada.



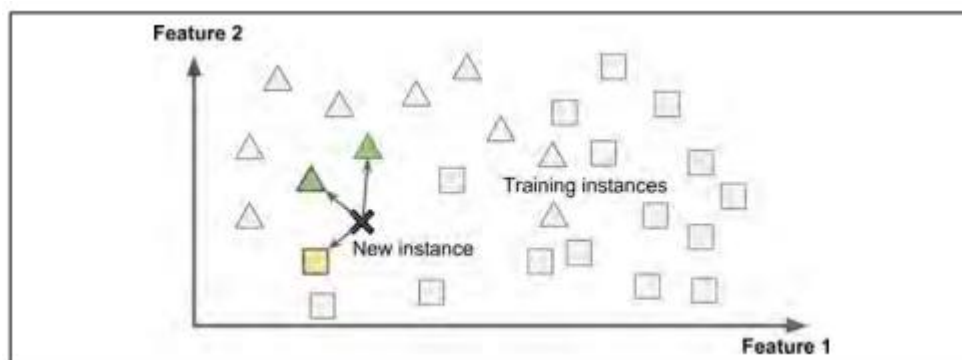
Fonte: SRIVASTAVA, 2015.

3.2.3.3 *Aprendizado baseado em Instância ou Modelo*

A última forma de classificar um Aprendizado de Máquinas é observando a forma como este cria generalizações, ou seja, depois de um número de treinamentos, o sistema precisa ser capaz de generalizar em dados novos.

Uma aprendizagem baseada em Instância é aquela em que o algoritmo mede a similaridade dos novos dados com o conjunto de treinamento. No caso do exemplo do filtro de spam em caixas de e-mail, o sistema identifica atributos como número de palavras em comum em e-mails que foram classificados como suspeitos, de modo que quanto mais próximo desse valor, maior será a probabilidade do novo e-mail ser de fato um spam. A Figura 16 mostra o uso da similaridade como base da aprendizagem baseada em instâncias.

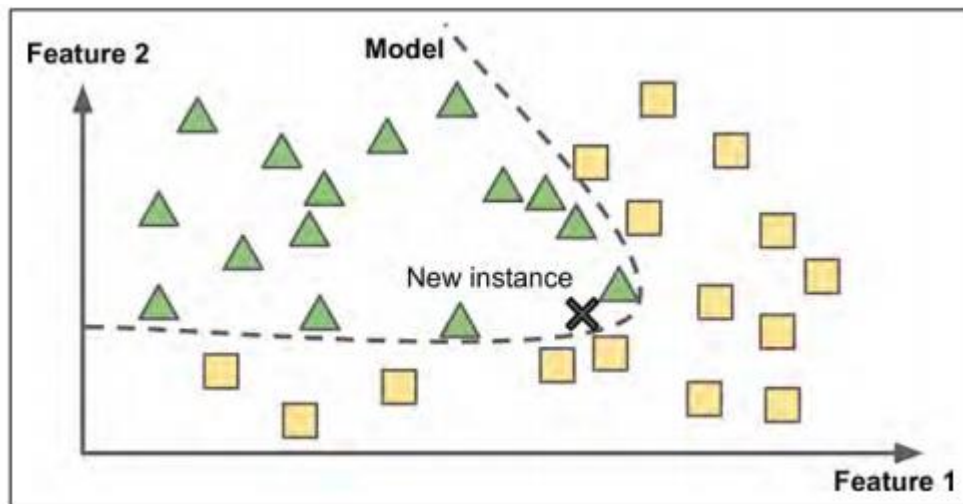
Figura 16: Aprendizado baseado em instâncias.



Fonte: GÉRON, 2017.

Já o aprendizado baseado em modelo, ilustrado na Figura 17, é aquele em que todas as suposições sobre o domínio do problema podem ser explícitas por uma equação matemática que representa o sistema. Neste tipo são utilizadas as regressões.

Figura 17: Aprendizado baseado em modelo.

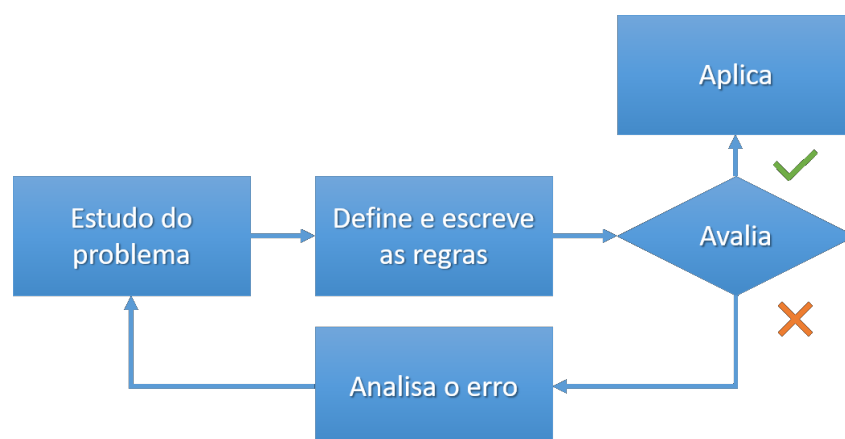


Fonte: GÉRON, 2017.

3.2.4 Composição básica de um algoritmo de Aprendizado de Máquinas

Ao observar os algoritmos tradicionais, é possível concluir que são compostos por um conjunto de regras, capazes de generalizar uma grande gama de possibilidades de atributos para os dados do problema a ser tratado, conforme ilustra a Figura 18 a seguir.

Figura 18: Esquema geral de um algoritmo tradicional.

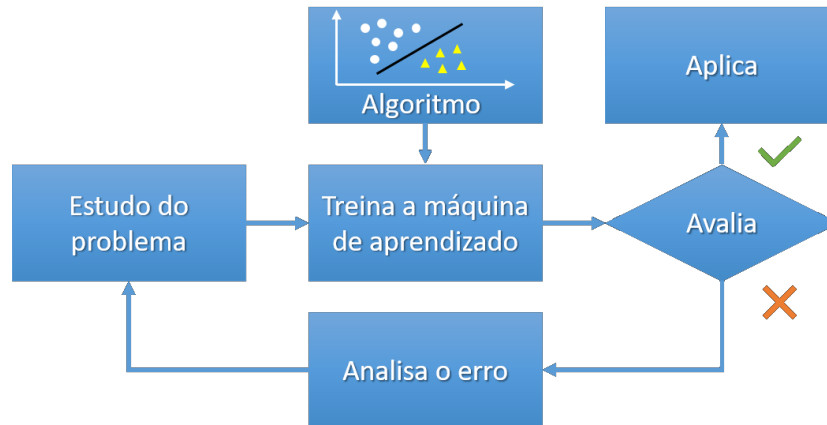


Fonte: Adaptado de GÉRON, 2017.

No entanto, se o problema for complexo, pode-se utilizar de dados históricos para entendê-lo melhor. Estudar o problema e aplicar todas as regras viáveis para a resolução do problema pode ser uma tarefa complexa, quando não impossível. Desta

maneira, a abordagem de um aprendizado de Máquinas se faz necessário. Aqui, o foco será estudar os modelos que melhor se encaixam para a resolução do problema, testar e analisar os erros, melhorando a abordagem a cada nova interação. A Figura 19 resume este esquema.

Figura 19: Esquema geral de um algoritmo de Machine Learning.

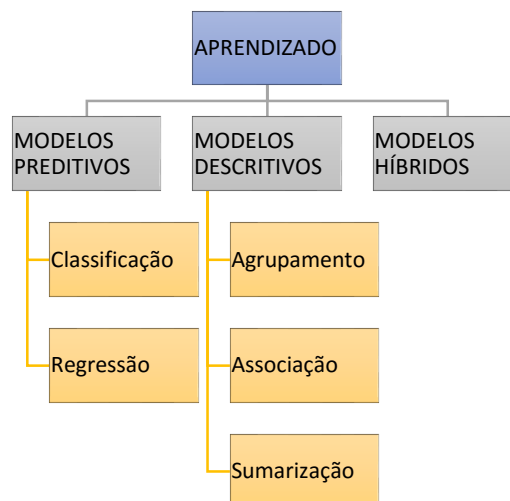


Fonte: Adaptado de GÉRON, 2017.

3.2.5 Modelos de Aprendizado de Máquinas

Quando se trata de modelos de Aprendizado de Máquinas, existem três principais: Aprendizagem Preditiva, Descritiva ou Híbrida. A Figura 20 a seguir ilustra essa divisão.

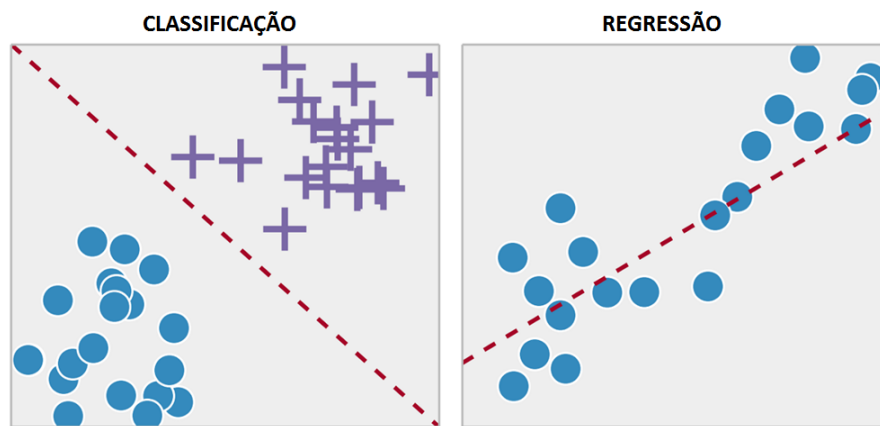
Figura 20: Modelos de Aprendizado de Máquinas.



Fonte: Autor, 2019.

A análise preditiva é indicada para prever comportamentos e resultados. Com base em dados históricos, o algoritmo trabalha buscando padrões e variações com o objetivo de determinar um comportamento ou classe. Na aprendizagem preditiva o programa é capaz de induzir um conceito a partir de um conjunto previamente conhecido e rotulado com suas respectivas classes. Se este rótulo é um número real, trata-se de uma regressão. Mas, em contrapartida, se este rótulo é proveniente de um conjunto finito e não ordenado, então é o caso de uma classificação. A diferença entre estes dois modelos é ilustrada na Figura 21 a seguir.

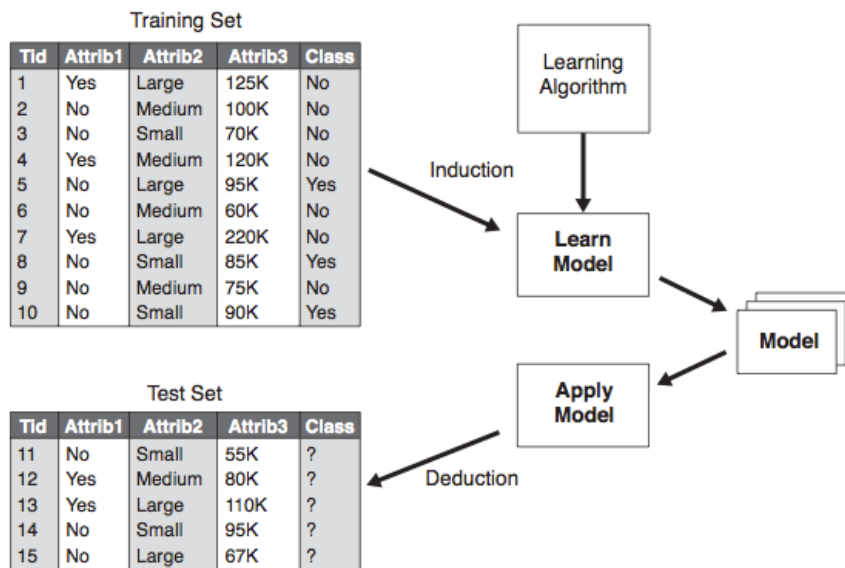
Figura 21: Comparativo entre modelos de classificação e regressão.



Fonte: SILVA, 2018.

Conforme será exposto e explicado no capítulo de desenvolvimento deste trabalho, o modelo de classificação foi definido como ideal para a solução proposta no projeto. A Figura 22 mostra uma abordagem genérica para a construção de classificadores.

Figura 22: Aprendizagem preditiva.



Fonte: TAN, STEINBACH, KARPATNE & KUMAR, 2005.

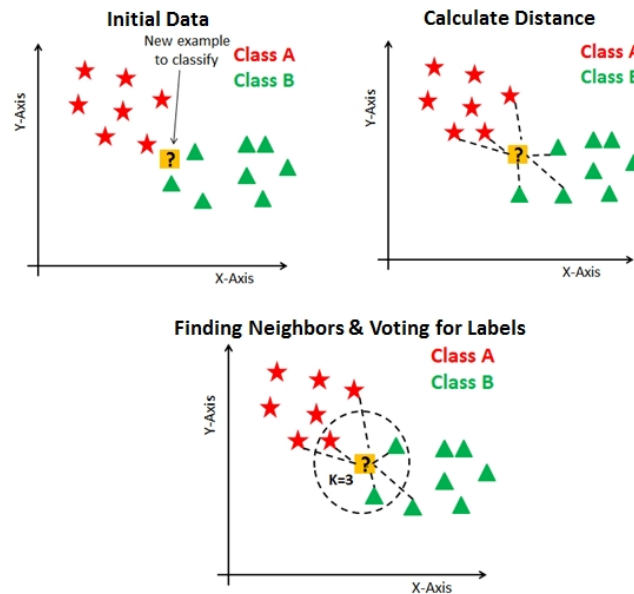
Já a análise descritiva busca a compreensão de dados do momento atual, trabalhando com valores históricos para gerar um panorama do que está acontecendo no tempo corrente. Trata-se de uma ótima maneira de visualizar e agrupar dados. Por não contemplar o escopo deste trabalho, não será detalhado este modelo.

3.2.5.1 Classificação K-NN

O modelo de classificação K-Nearest Neighbors (K-NN) foi um dos primeiros a serem desenvolvidos na área de Aprendizado de Máquinas. Apesar de requerer grande esforço computacional, é um dos mais utilizados para reconhecimento de padrões. Trata-se de um modelo baseado em instâncias, que utiliza do critério de similaridade por medidas de distância no espaço multidimensional dos dados, usualmente a distância Euclidiana.

O modelo de treinamento contém todos os parâmetros e seus respectivos rótulos. Quando novos dados são carregados, são calculadas as distâncias entre as observações, para que as mais próximas sejam selecionadas, e a classe determinada. Na Figura 23 é possível visualizar essa dinâmica: um novo exemplo a ser classificado é inserido, as distâncias das observações são estimadas e, por meio de cálculos e da definição do número de vizinhos, a classe é determinada.

Figura 23: Exemplo de funcionamento do modelo K-NN.



Fonte: NALANI, 2018.

A distância Euclidiana é definida como:

$$Distance(X_1, X_2) = \sqrt{\sum_{i=1}^k (X_{1i} - X_{2i})^2} \quad (1)$$

Onde X_1 e X_2 são vetores de atributos.

Na classificação pelo modelo K-NN a classe pode ser determinada pela moda do atributo-classe para os K vizinhos selecionados. O valor de K pode ser determinado experimentalmente, de modo que quanto maior for o conjunto de treinamento, maior será o valor de K.

Trata-se de um modelo caro computacionalmente, pois exigirá uma boa capacidade de processamento para sistemas com grande volume de dados. Outra desvantagem é que sua velocidade pode ser lenta, uma vez que todas as distâncias necessárias devem ser calculadas. No entanto, o K-NN supera uma grande variedade de modelos pela sua acurácia, por se tratar de um modelo de reconhecimento de padrões bastante poderoso.

3.2.6 Desafios na Implementação de Algoritmos de Aprendizado de Máquinas

A implementação de algoritmos de Machine Learning requer uma série de cuidados, para que não haja desperdício de tempo e recursos em estratégias ineficientes que não agregam nenhum tipo de valor ao negócio. A seguir são apresentados os principais riscos à efetividade das soluções em aprendizado de máquinas:

- **Quantidade Insuficiente de dados para treinamento do algoritmo:** caso não sejam fornecidos grandes volumes de dados, não haverá reconhecimento de padrões suficiente para determinação de classes e rótulos nos dados futuros. Até mesmo problemas simples exigem grandes quantidades de dados;
- **Dados não representativos:** o algoritmo deve ser capaz de criar generalizações, de modo que se for alimentado com dados que não representa o conjunto, essa capacidade será arruinada. Os dados para treinamento do algoritmo devem ser amplos e representar uma boa parte do universo de possibilidades para o problema representado;
- **Dados sem qualidade:** se o conjunto de dados estiver repleto de outliers, ruídos, valores nulos ou zerados, a capacidade de aprendizado do algoritmo será comprometida. Por isso é tão importante a etapa de limpeza e extração dos dados;
- **Variáveis irrelevantes:** o uso demasiado de variáveis em modelos preditivos e descritivos atrapalha o desempenho de aprendizado do algoritmo. Deve-se selecionar apenas as variáveis que realmente influenciam no processo, trabalhando para melhorar a sua usabilidade com a descoberta de novas variáveis, e não incrementando todas ao dataset;
- **Overfitting dos dados de treinamento:** se o modelo for muito complexo para a quantidade de dados efetivos e ruidosos, a generalização será tão grande que não irá representar o sistema. Será necessário reduzir a quantidade de variáveis, adquirir mais dados, remover ruídos e outliers, ou seja, dados que se diferenciam drasticamente dos demais do conjunto;
- **Underfitting dos dados de treinamento:** de maneira oposta ao anterior, ocorre quando o modelo é simples demais para o entendimento da estrutura dos dados. Para resolver esse problema é necessário selecionar um modelo mais complexo, com mais parâmetros, selecionar variáveis com maior relação ao problema e reduzir as restrições do programa.

3.2.7 Validações e Testes

Escolhido o modelo a ser implementado e tomando os devidos cuidados para aplicação dos métodos, incluindo o pré-processamento dos dados, é preciso avaliar o desempenho do classificador desenvolvido. Uma forma de se mensurar é observando a taxa de erros, ou seja, o percentual de acertos da classe determinada pela quantidade de predições realizadas. Para realização desta validação experimental, pode-se utilizar um método de partição dos dados disponíveis em dois subconjuntos, sendo um de treinamento e o outro de teste. Esta maneira é denominada **Hold-out**, e tipicamente utiliza dois terços das observações como treinamento e o restante como teste, sendo as medidas de desempenho aplicadas neste último. Outra maneira de realizar essa divisão é a chamada **Validação Cruzada** (ou k-Fold Cross Validation), em que os dados são segregados em K subconjuntos, de tamanho aproximadamente iguais, de modo que treinamento e testes são realizados K vezes.

Para validação, utiliza-se os conceitos de acurácia, precisão e sensibilidade. Ela pode ser definida utilizando as seguintes medidas:

- Verdadeiro Positivo (VP): número de observações corretamente classificadas como positivo;
- Verdadeiro Negativo (VN): número de observações corretamente classificadas como negativo;
- Falso Positivo (FP): número de observações erroneamente classificadas como positivo;
- Falso Negativo (FN): número de observações erroneamente classificadas como negativo;
- N: número total de observações classificadas.

Sendo assim, a acurácia é expressa como:

$$Acur\acute{a}ci\ a = \frac{VP + VN}{N} \quad (2)$$

Já a precisão mede a taxa de acertos entre as observações classificadas como positivas, podendo ser definida como:

$$Precisão = \frac{VP}{VP + FP} \quad (3)$$

A sensibilidade, ou também chamada de Recall, é a taxa de acertos de observações positivas corretamente classificadas:

$$Sensibilidade = \frac{VP}{VP + FN} \quad (4)$$

O F1-Score busca uma média harmônica entre precisão e recall:

$$F1 - Score = \frac{2 * Precisão * Sensibilidade}{Precisão + Sensibilidade} \quad (5)$$

Uma opção viável é observar a Matriz de Confusão, que compara todos estes valores, conforme ilustra a Figura 24.

Figura 24: Matriz de Confusão.

		Valor Verdadeiro (confirmado por análise)	
		positivos	negativos
Valor Previsto (predito pelo teste)	positivos	VP Verdadeiro Positivo	FP Falso Positivo
	negativos	FN Falso Negativo	VN Verdadeiro Negativo

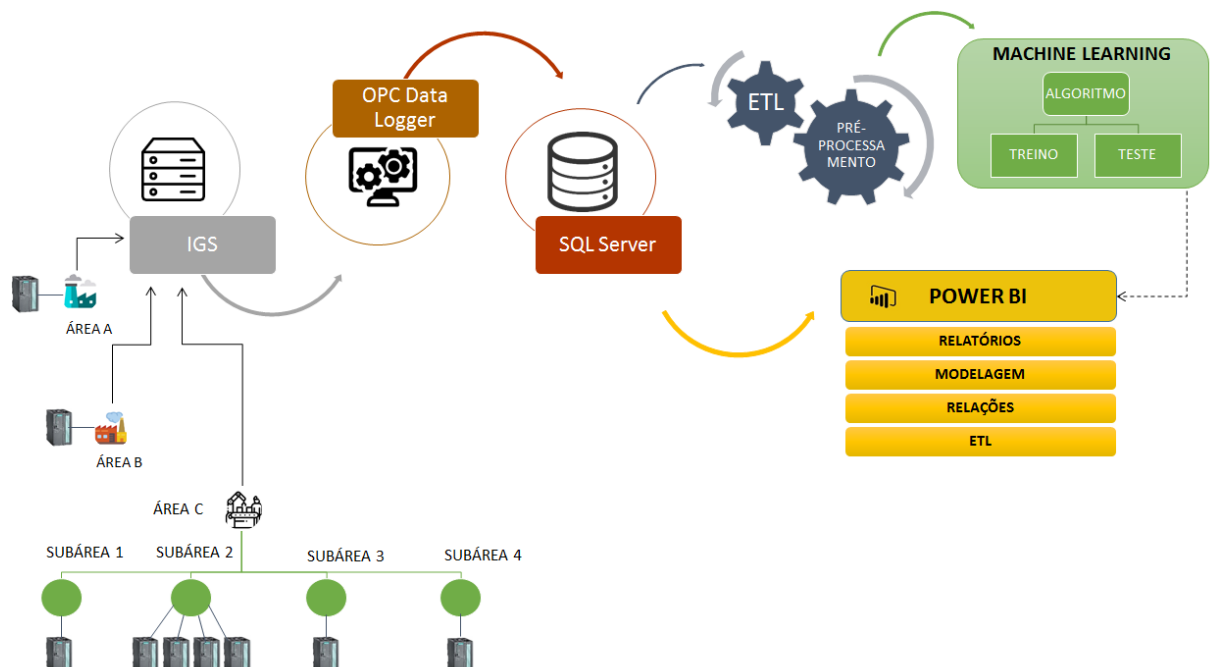
Fonte: SOUZA, 2009.

4 DESENVOLVIMENTO

Para este trabalho, propôs-se a construção de um sistema integrado, capaz de fornecer o monitoramento e a predição de indicadores de desempenho de um processo industrial. Desta maneira, o projeto foi dividido em duas grandes vertentes: uma ferramenta de BI e o estudo de modelos de predição capazes de gerar resultados de valor para o processo.

Desta maneira, a primeira etapa foi o desenvolvimento de uma ferramenta de BI capaz de fornecer o monitoramento e análise de informações importantes para tomada de decisão no processo operacional da área central de produção de uma indústria. Para este feito, foi necessário edificar toda a arquitetura de um programa de Business Intelligence, passando desde a coleta, limpeza e transformação de dados, até o front-end dos dashboards e relatórios destinados à gerência da organização, incluindo avaliação e melhoria do projeto. O esquema geral da ferramenta é exposto na Figura 25 a seguir.

Figura 25: Esquema geral.



Fonte: Autor, 2019.

4.1 Coleta de Dados

A primeira e mais importante etapa do ciclo do BI é a coleta de dados. Nesta fase, é importante selecionar dados que permitirão construir indicadores que serão apresentados na interface final do projeto. A seguir apresenta-se o desenvolvimento da estrutura e recursos utilizados para tal.

4.1.1 Dados de Processo

A área produtiva escolhida para o desenvolvimento da ferramenta de BI possui estrutura de automação já consolidada e nenhuma alteração foi realizada. Toda a instrumentação e controle da planta foi mantida, nenhum sensor adicionado, de modo que os dados a serem coletados derivam de variáveis de processo já consolidados. O entendimento dessa estrutura é importante para a realização da coleta.

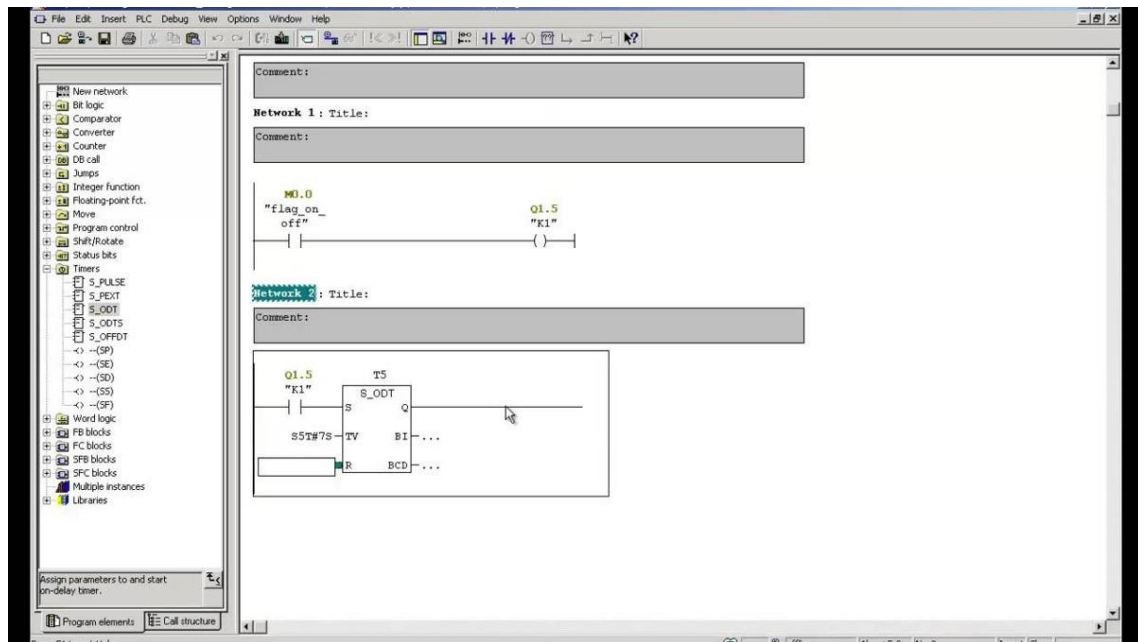
O controle do processo é realizado pelos Controladores Lógico Programáveis (CLP's, ou PLC's, da sigla em inglês) do fabricante Siemens, modelos S7-300 e S7-400. O ambiente de interação com os PLC's denomina-se Step 7, do mesmo fabricante, e oferece todo o suporte e confiança para configuração, programação, updates, comissionamento e operação.

Figura 26:- PLCs S7-300 e S7-400 Siemens



Fonte: Siemens, 2019.

Figura 27: Ambiente do Step 7.

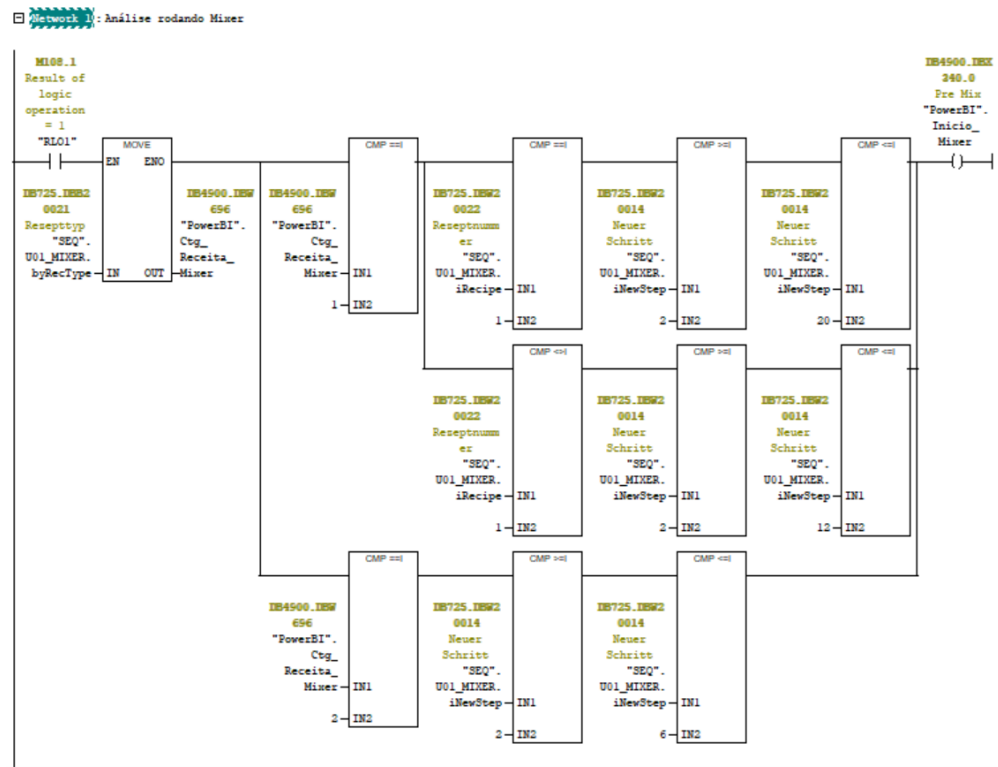


Fonte: Siemens, 2019.

Para realizar a coleta dos dados, tornou-se necessário a preparação de dados no ambiente do Step 7, com objetivos de:

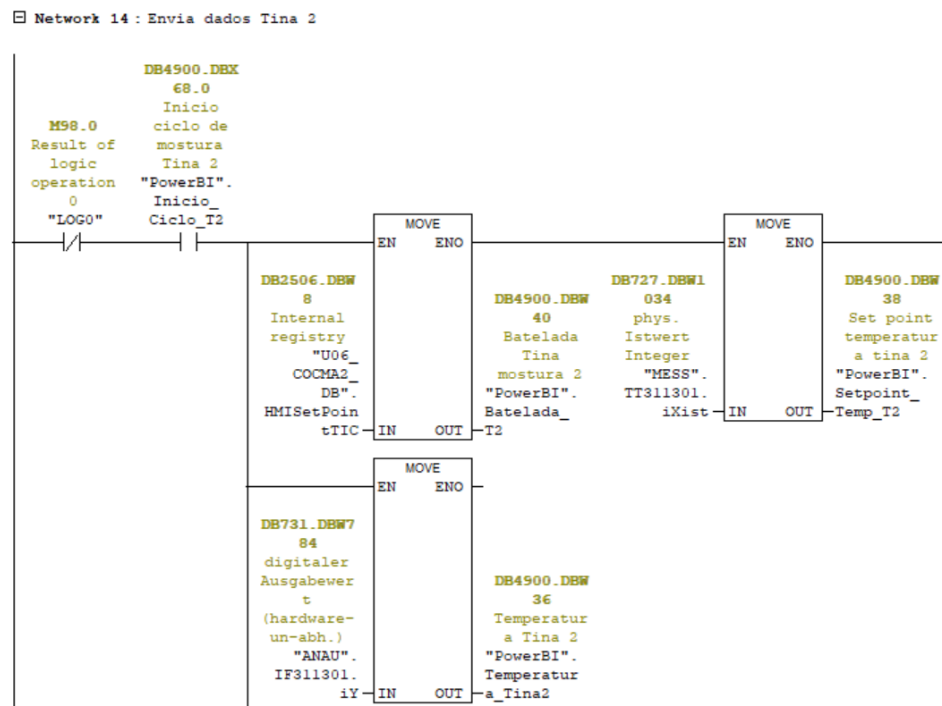
- Coletar dados somente no período de funcionamento efetivo dos equipamentos durante o processo produtivo. Para isso foi necessário analisar as receitas de cada produto, entendendo em qual passo em cada uma delas seria iniciado a coleta, conforme exemplo da Figura 28;
- Organizar dados em um único grupo de variáveis dedicado para o projeto, com a finalidade de facilitar o trabalho e garantir a não interferência no processo e funcionamento dos equipamentos entendendo em qual passo em cada uma delas seria iniciado a coleta. A solução para isso foi criar uma FC (Function) para cada PLC de processo dedicado ao projeto, movendo as variáveis de processo escolhidas, conforme exemplo da Figura 29.

Figura 28: Exemplo de lógica de análise de equipamentos em funcionamento.



Fonte: Autor, 2019

Figura 29: Exemplo de lógica para mover variáveis no Step 7.



Fonte: Autor, 2019

Sendo assim, três grandes áreas produtivas foram envolvidas, sendo a área C a de maior importância para este trabalho. Um total de variáveis foram configuradas em 09 PLC's. A Tabela a seguir lista os PLC's por área e a quantidade de variáveis coletadas em cada um deles.

Tabela 6: Relação da quantidade de PLCs e variáveis coletadas por área.

Área	Subárea	PLC	Quantidade de Variáveis
A	-	PLC_A100	6
B	-	PLC_B100	4
C	1	PLC_C100	456
	2	PLC_C200	335
		PLC_C201	520
		PLC_C202	40
	3	PLC_C300	30
	4	PLC_C400	106

Fonte: Autor, 2019.

4.1.2 OPC Server – IGS

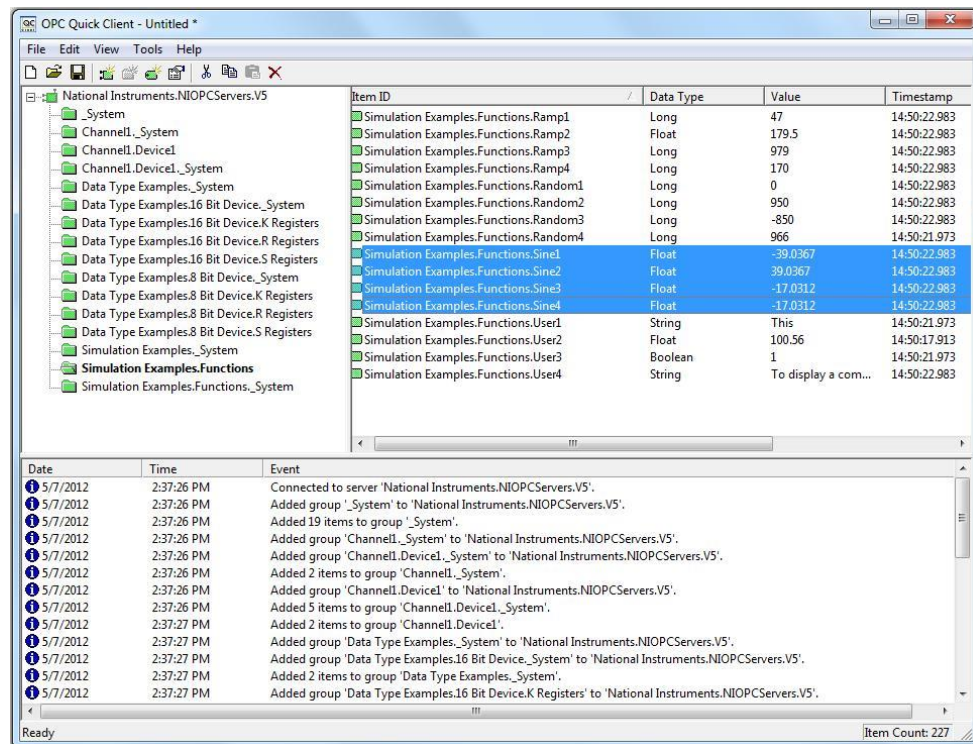
OPC (OLE for Process Control) é um conjunto de padrões de comunicação de dados para a indústria que utiliza a tecnologia OLE (Object Linking and Embedding), da fabricante Microsoft para sistemas Windows, capaz de conectar objetos de dados. Algumas características do OPC são:

- Troca dados utilizando protocolo TCP (Transmission Control Protocol) otimizado;
- Permite configurações de segurança;
- Possui arquitetura robusta, garantindo comunicação confiável e detecção automática de falhas;

Desta forma, uma solução OPC da GE (General Electric) foi escolhida para realizar a coleta dos dados de processo. O software é denominado IGS (Industrial Gateway Server) e é um servidor OPC UA (Unified Architecture), capaz de oferecer conectividade confiável e robusta, além de ter uma interface fácil de ser utilizada e configurada. O OPC UA é um protocolo para comunicação industrial, padronizado na IEC 62541, aplicável a diversos componentes, em todos os domínios industriais, visando a troca de informações entre sensores, atuadores, sistemas de controle,

sistemas de fabricação e sistemas de planejamento. Para comunicação com os PLC's Siemens, o IGS utiliza drivers padrão Profinet.

Figura 30: RunTime do IGS.

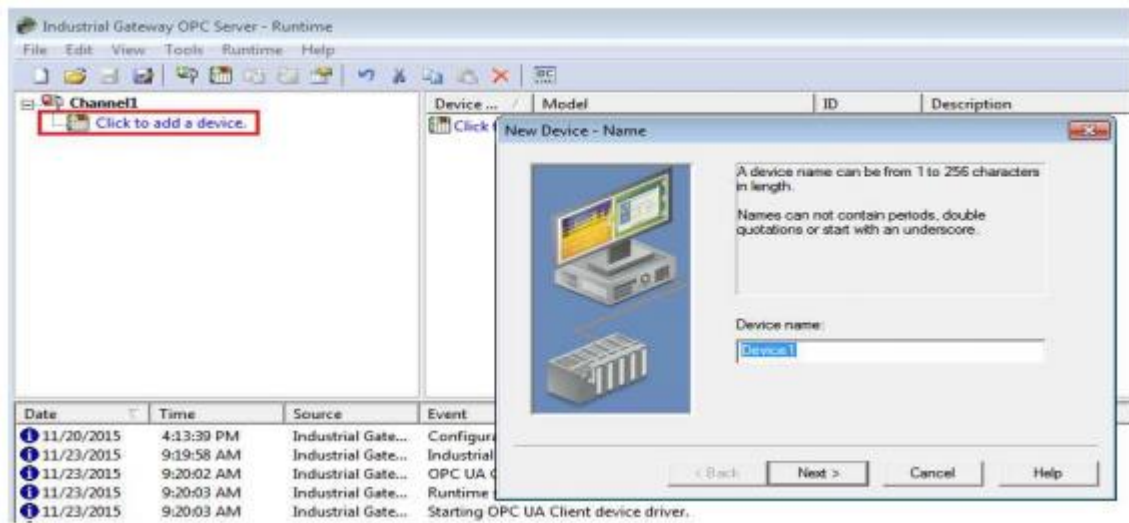


Fonte: GE, 2019.

O IGS tem a capacidade de se conectar a vários PLC's, organizando os dados coletados em grupos, o que foi fundamental, uma vez que as variáveis escolhidas derivam de várias partes diferentes do processo produtivo. Além disso, permite conectar-se em RunTime, visualizar os dados em tempo real, certificando-se de que o dado coletado está correto e de acordo com o esperado.

Para realizar a comunicação, o IGS utiliza-se de drivers padrões dos protocolos de comunicação. No caso da Siemens, o driver utilizado é o Siemens TCP/IP Ethernet. O software tem a capacidade de se conectar com dispositivos utilizando simultaneamente até 80 diferentes drivers. Os dispositivos devem ser configurados para conexão e ficam disponibilizados em canais, conforme mostra a Figura 31.

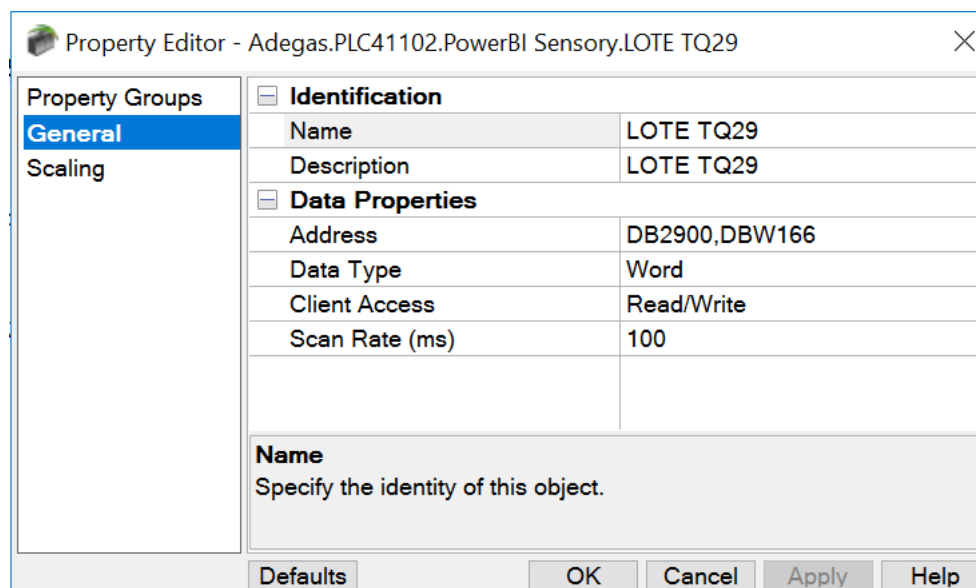
Figura 31: Configuração de dispositivos no IGS.



Fonte: GE, 2019.

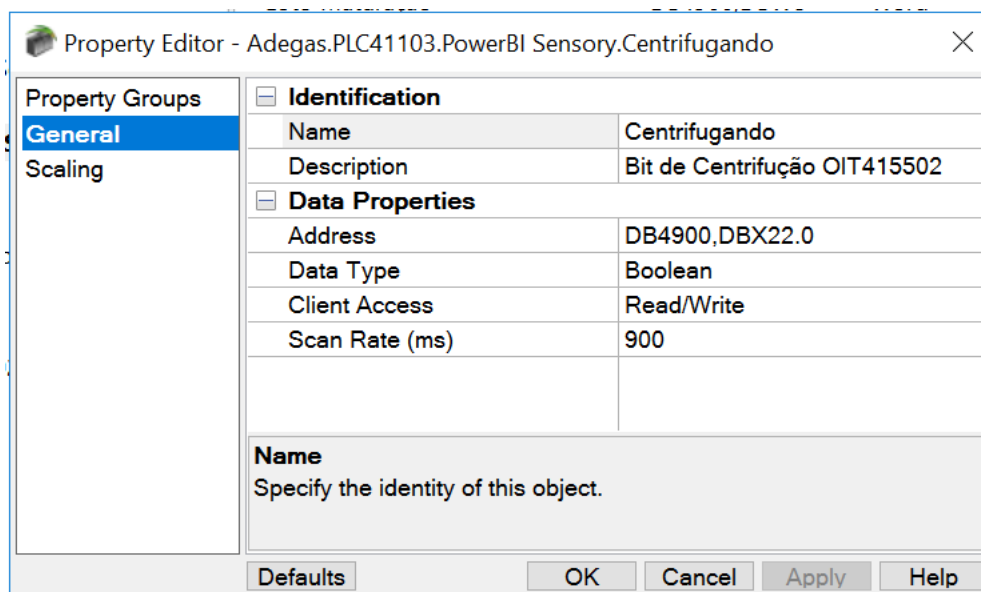
A configuração de uma variável no ambiente do IGS é feita a partir da conexão com o PLC escolhido e adicionando uma nova TAG, definindo um nome e descrição para o dado. A partir disso, associa-se o endereço da variável no PLC, o tipo (Word, DoubleWord, Int, Float, Boolean, etc), seu modo (escrita e ou leitura) e sua taxa de coleta. A Figura 32 mostra um exemplo de uma variável do tipo Word, enquanto a Figura 33 mostra uma variável booleana, com suas configurações no ambiente do IGS.

Figura 32: Configuração de uma variável do tipo Word no IGS.



Fonte: Autor, 2019.

Figura 33: Configuração de uma variável booleana no IGS.



Fonte: Autor, 2019.

4.1.3 OPC Client – Advanced OPC Data Logger

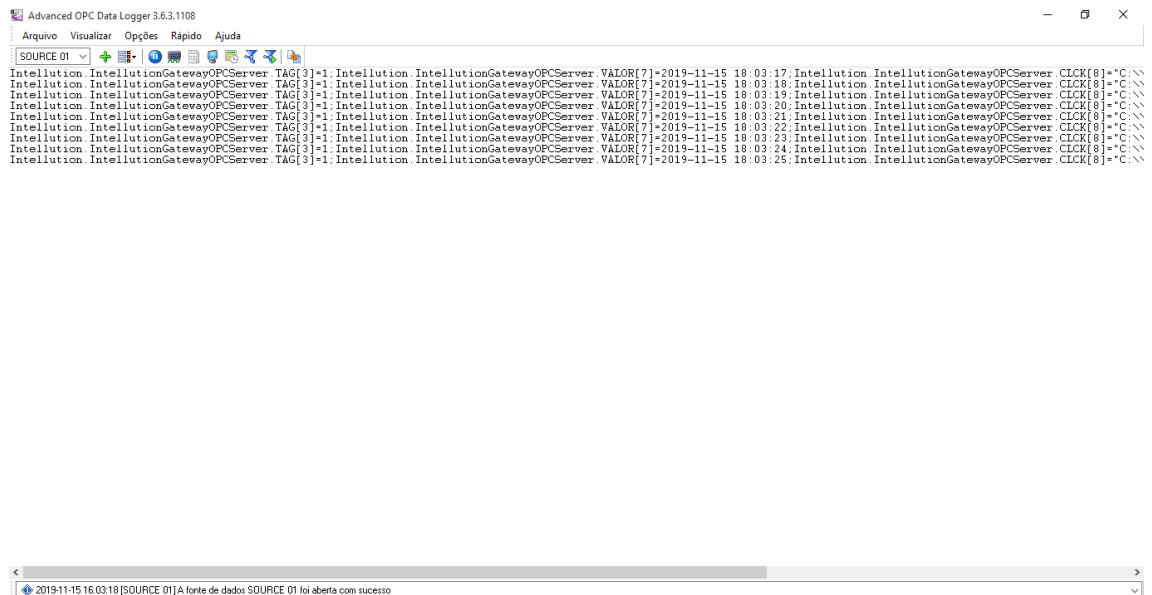
Enquanto o OPC Server estabelece uma conexão com o processo, é necessária uma ferramenta capaz de se conectar com os recursos onde serão alocados e armazenados os dados. Para tal, foi escolhido a solução da AGG Software, o Advanced OPC Data Logger.

O software foi projetado para enviar e receber dados para um servidor OPC. Quando recebidos, os dados são processados e agrupados de acordo com as configurações de usuário, e depois são exportados para uma aplicação conveniente. É capaz de enviar dados para arquivos de texto, arquivos binários, planilhas, Access, SQL Server, MySQL, MSSQL, bancos de dados compatíveis com o Oracle ou ODBC, e outros destinos.

As principais razões para a escolha dessa solução foram seu custo reduzido, aliado à simplicidade de uso do software, e a facilidade de conexão com o mecanismo de Banco de Dados.

A primeira etapa na configuração do OPC Data Logger é a conexão com o servidor. Posteriormente, é necessário adicionar as variáveis desejadas, com seus respectivos endereços e configurações de tipo. Os dados serão exibidos na tela na frequência escolhida, como mostra a figura a seguir.

Figura 34: Ambiente do OPC Data Logger



Fonte: Autor, 2019.

Uma configuração importante é a taxa de coleta dos dados. No caso do IGS vimos que é possível inserir altas taxas de coleta, da ordem de 100ms. Para determinar esse valor é preciso analisar o processo em questão. Perguntas importantes para auxiliar nessa decisão são:

- Com que frequência os dados variam?
- Caso uma ação de processo seja tomada, quanto tempo essa variável em questão levará para sofrer uma variação significativa?
- O que é considerado significativo de mudança para as análises em questão?

Dessa maneira, para cada processo a ser analisado foi escolhido uma taxa conveniente. Casos como valores de oxigênio incorporado ao líquido em uma linha de envase de garrafas e latas, com alta rotatividade e velocidade, exigiriam uma taxa muito alta. Porém, com a observação em campo, constatou-se que a variação é muito mais devido ao fluxo nas tubulações da linha. Isso permitiu a escolha de uma taxa de 5s como suficiente. Em outro caso, pequenas variações de temperatura em um tanque são percebidas e importantes para o processo e, por isso, devem ser percebidas nas análises. Isso demanda um menor tempo entre uma coleta e outra. Além disso, o dado pode ser compactado, ou seja, pode ser definido um percentual de variação do valor para que a coleta ocorra. Essa opção garante menor utilização de memória e evita

dados desnecessários, porém deve ser avaliada conforme as necessidades das ferramentas que irão consumir os dados. Para algumas soluções, a visualização de todo o conjunto de dados ao longo de um período pode ser necessária. A Figura 35 ilustra uma configuração no OPC Data Logger, utilizando compactação com um dead band de 0,1%.

Figura 35: Configuração de coleta com compactação de dados.

Group properties

Name: Brassagem

Update rate (msec): 1000

Dead band (%): 0,1

☒ Active
☐ Wait till all items are updated
☐ Add server name to item name
☐ Export whole group
☒ Export one row for all changed items
☐ Export at fixed interval

☒ Export valid values only
☒ Export date/time stamp
☐ Timestamps in local time
☐ Export item quality
☒ Export item name
☒ Don't export NULL values

Export at

☒ Use update rate
☐ Time of day
☐ Time using Unix Cron schedule

11:00:12:00

WD: * M: * D: * ?
 H: * MIN: */1 SEC: 0

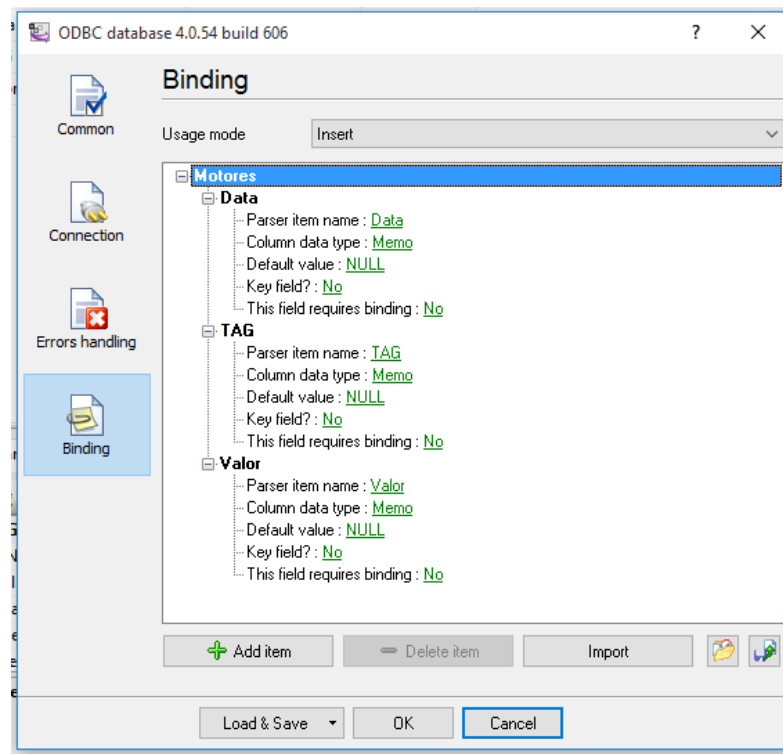
OK Cancel

Fonte: Autor, 2019.

O próximo passo foi estabelecer uma fonte de dados ODBC capaz de se conectar com a instância do Banco de Dados, no caso do SQL Server. Este recurso é acessado no Administrador de Fonte de Dados ODBC do Windows. Uma nova fonte é adicionada, selecionando o driver para SQL Server e, após todas as configurações, um teste de conexão é realizado, indicando que foi estabelecido com sucesso.

Com a fonte ODBC configurada, é possível configurar o OPC Data Logger para enviar os dados para serem armazenados no SQL Server, conforme a figura a seguir.

Figura 36: Conexão do OPC Data Logger com tabelas do SQL Server.



Fonte: Autor, 2019.

4.2 Banco de Dados – SQL Server

Conforme exposto no capítulo sobre Business Intelligence, um recurso de armazenamento dos dados é essencial para consolidação de uma ferramenta de BI. Os dados operacionais devem ser armazenados para posteriormente serem tratados e carregados nas análises de usuário. Para a construção da ferramenta desse projeto, optou-se pelo banco de dados da Microsoft, o MSSQL. Trata-se de um sistema gerenciador de banco de dados relacional, cuja linguagem de consulta primária é o Transact SQL. É uma solução capaz de atender às demandas desde os mais simples negócios até os mais complexos cenários. Os principais recursos do SQL Server são listados a seguir:

- Database Engine: é o serviço de armazenamento, proteção e processamento dos dados. Fornece acesso controlado e processamento rápido de transações.
- Management Studio: usado para configurar, gerenciar e administrar todos os recursos do SQL Server

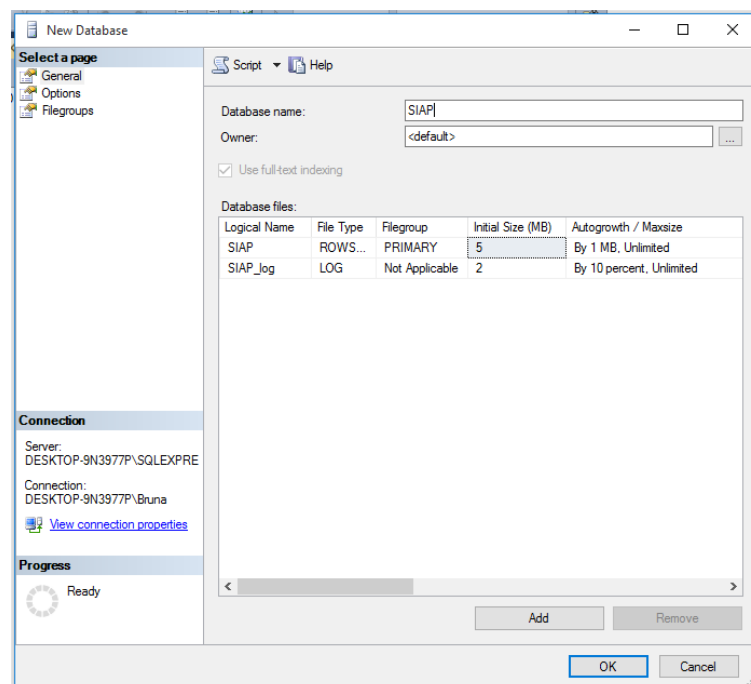
- Reporting Services: fornece um conjunto de ferramentas e serviços locais para criar, implantar e gerenciar relatórios paginados e móveis.
- Analysis Services: dá suporte a modelos de tabela em todos os níveis de compatibilidade, utilizado como suporte para tomadas de decisão
- Integration Services: possibilita limpeza, seleção e transformação de dados, por meio de tarefas internas.

SQL Sever 2016 Standard foi a versão escolhida para o projeto, uma vez que a organização já possuía licença do banco de dados disponível para ser utilizada.

Depois que as configurações de ODBC foram estabelecidas, é possível acessar a instância padrão de banco de dados do SQL ou criar uma nova. Desta maneira, no SQL Management Studio é possível conectar-se ao Servidor da instância em questão. Essa conexão pode ser local ou remota, utilizando credenciais próprias do banco de dados, recurso bastante útil para acessar o banco de dados em qualquer servidor de uma rede industrial.

Após conectado à uma instância do SQL, é possível criar o Banco de Dados, onde serão armazenadas as tabelas. Aqui a opção foi configurar todos os dados dentro do mesmo Banco de Dados, para facilitar os backups a serem realizados. No entanto, pode-se criar bancos separados por área, ou assunto caso seja necessário.

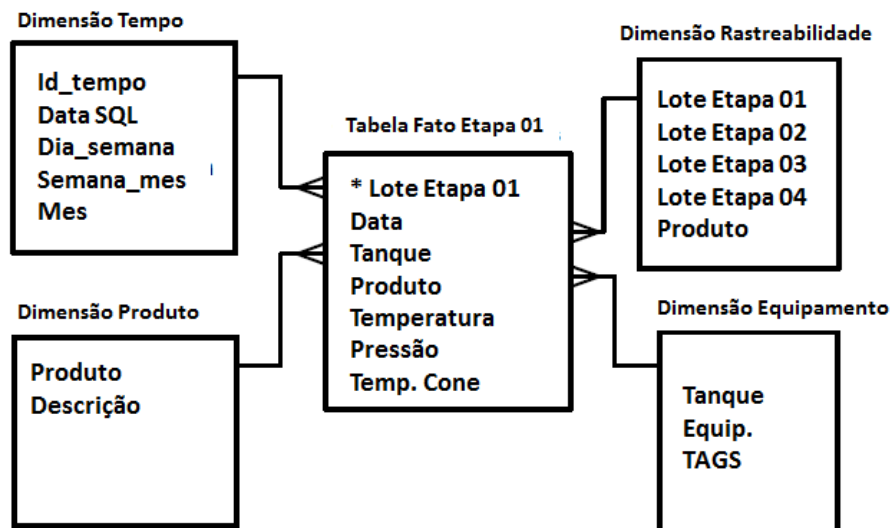
Figura 37: Novo Banco de Dados no SQL.



Fonte: Autor, 2019

Para criação das tabelas contendo os dados operacionais deste trabalho, foi necessário realizar um levantamento de todas as informações necessárias e como elas se relacionam, para garantir uma boa estrutura do banco de dados. O modelo utilizado foi o modelo denominado estrela, conforme o exemplo da Figura 38 a seguir.

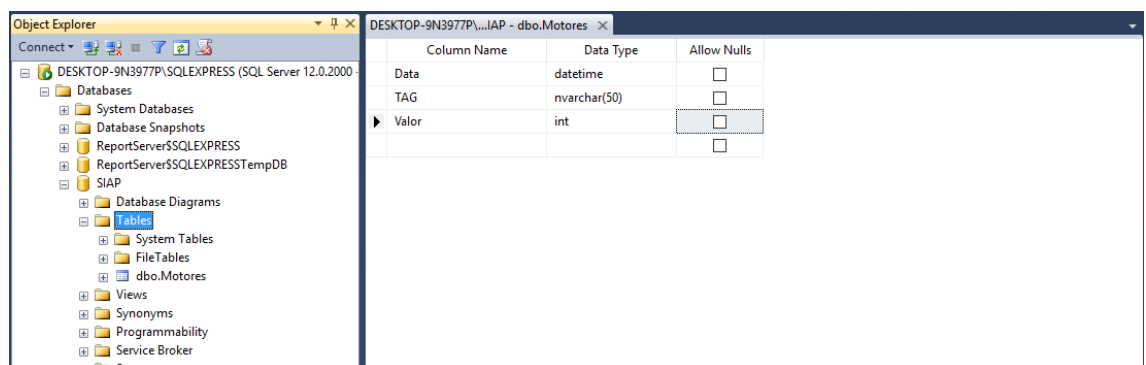
Figura 38: Exemplo de modelagem utilizada no Banco de Dados.



Fonte: Autor, 2019

Para criar uma nova tabela, dentro do banco de dados, basta informar o nome da coluna e o tipo de dados que cada uma delas deverá receber. Além disso é possível configurar se a tabela recebe valores nulos ou não, casos que ocorrem quando o PLC não reconhece o valor da variável de campo, saturação do instrumento de medição, erros de comunicação, etc. A Figura 39 mostra essas configurações da tabela Motores criada no banco de dados.

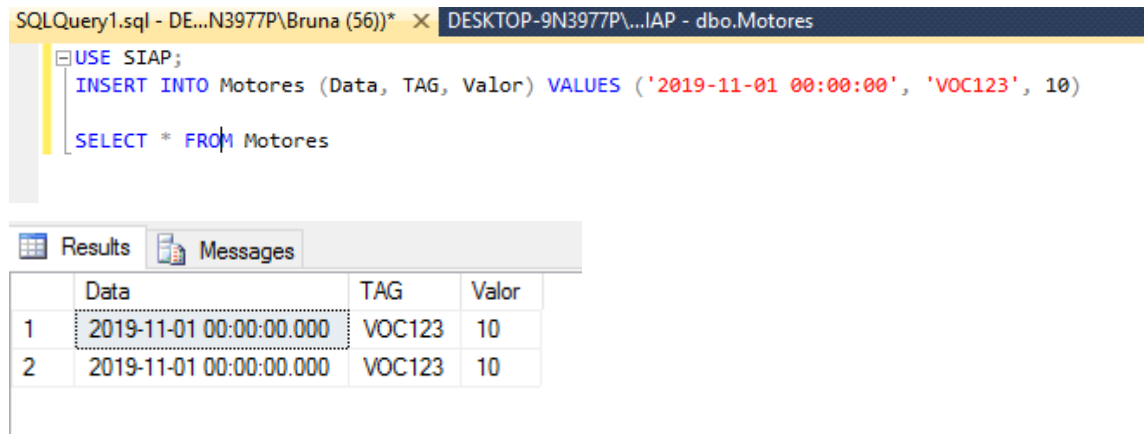
Figura 39: Nova tabela de dados no SQL.



Fonte: Autor, 2019.

Através da guia ferramenta de Query do SQL é possível realizar as operações básicas CRUD (Create, Read, Update, Delete) de um sistema de banco de dados. A Figura 40 mostra um exemplo de Query e a tabela com seus valores inseridos.

Figura 40: Operações básicas no Query do SQL.



SQLQuery1.sql - DE...N3977P\Bruna (56))* X DESKTOP-9N3977P\...IAP - dbo.Motores

```
USE SIAP;
INSERT INTO Motores (Data, TAG, Valor) VALUES ('2019-11-01 00:00:00', 'VOC123', 10)
SELECT * FROM Motores
```

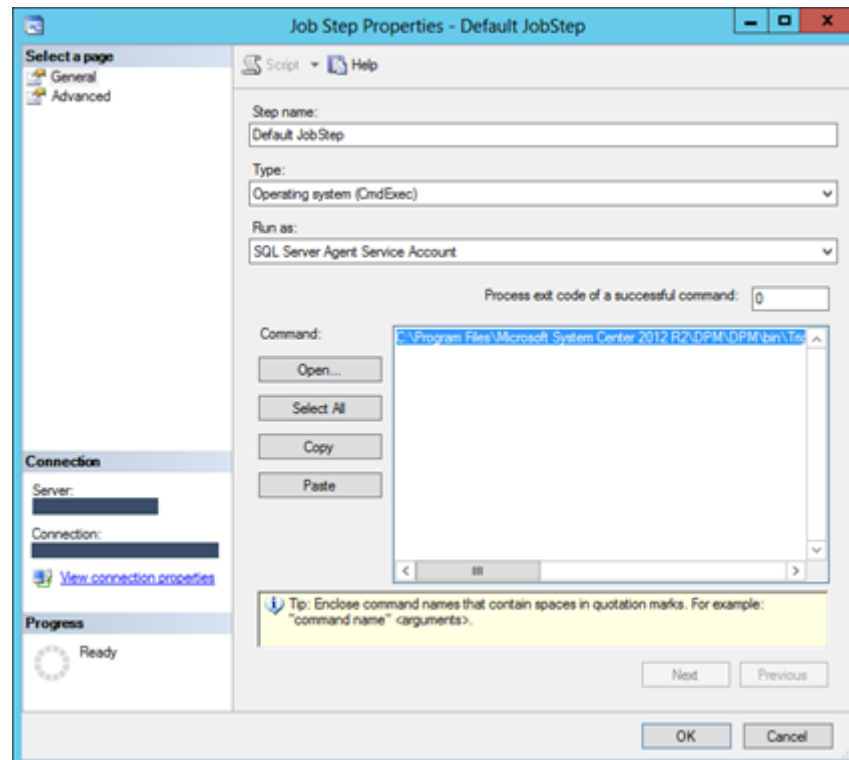
	Data	TAG	Valor
1	2019-11-01 00:00:00.000	VOC123	10
2	2019-11-01 00:00:00.000	VOC123	10

Fonte: Autor, 2019.

Assim, todas as configurações de coleta de dados são estabelecidas e podem ser realizadas consultas, leituras, atualizações ou mesmo deletar informações do banco de dados, via comandos SQL.

Outra funcionalidade importante é o SQL Server Agent, que permite que sejam agendadas tarefas no banco de dados. Com este recurso foi possível o agendamento de funções como realizar uma pré-limpeza dos dados, efetuar backups, aplicar tratamento prévio em alguns dados, etc. O agendamento especifica quando o trabalho deve ser executado, com data e horário específicos. A Figura 41 mostra a tela de configuração desse agendamento. As instruções são realizadas em linguagem Transact-SQL.

Figura 41: Configuração de Tarefas agendadas no SQL Agent.



Fonte: Autor, 2019.

4.3 Dashboards de Monitoramento e Análises no Power BI

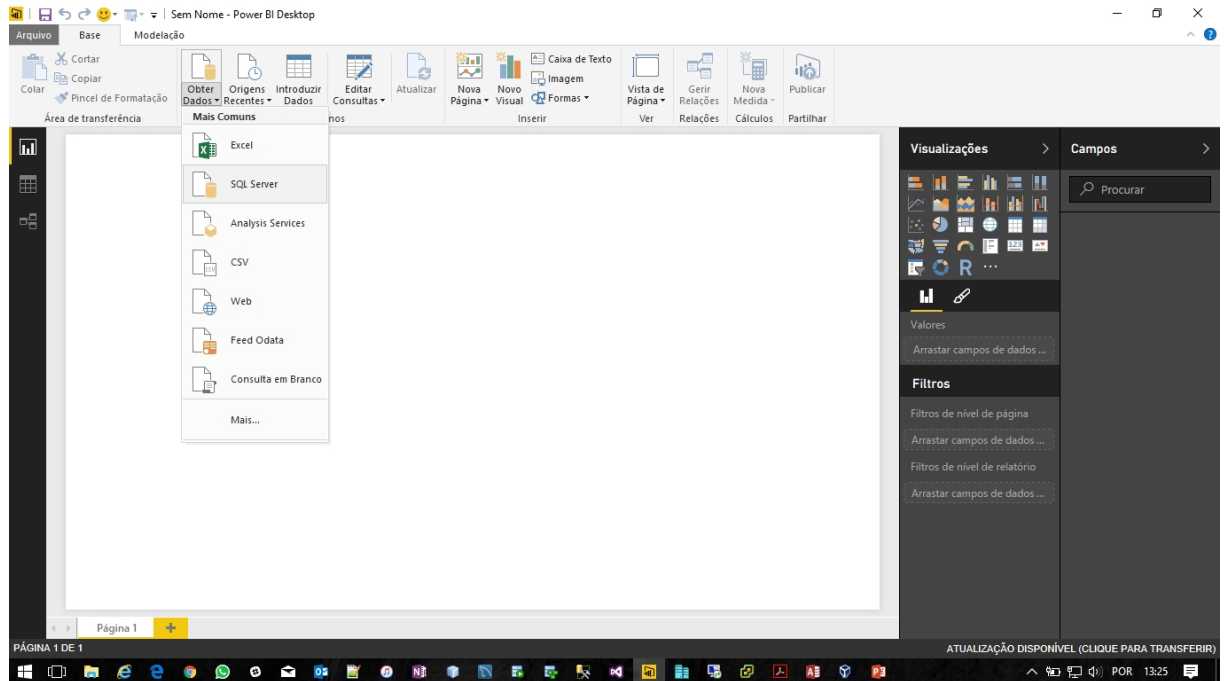
O Power BI, conforme já enunciado anteriormente, é líder de mercado nas soluções em BI e foi escolhido por esta razão para o desenvolvimento deste trabalho. Além disso, dois fatores foram cruciais nesta decisão: o custo reduzido de licenciamento, fator importante para iniciativas piloto em uma organização, e a facilidade de uso, tanto para os usuários finais, quanto para os desenvolvedores.

Suas principais funcionalidades são descritas a seguir.

4.3.1.1 Conexão e configuração de fontes de Dados

O Power BI permite conectar-se com fontes de dados operacionais diversas, como: Access, Excel, SQL Server, My SQL, XML, SAP, Facebook, e muitos outros. Neste projeto, foram utilizados o SQL Server e planilhas Excel como fonte de dados. A Figura 42 mostra a guia de inserção de novas fontes de dados.

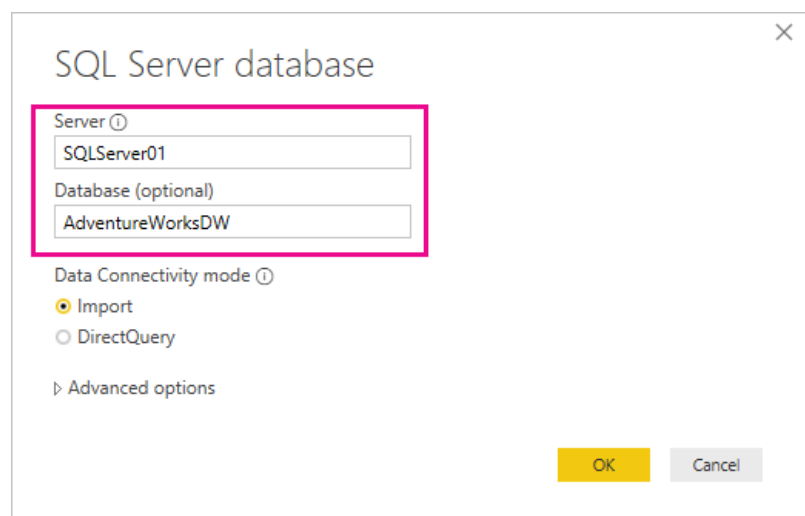
Figura 42: Conexão do Power BI com a fonte de dados.



Fonte: Microsoft, 2019.

A conexão poderá ser via Importação, com o dataset interno ao Power BI, ou DirectQuery, o qual os dados ficam no banco de dados de origem, e conforme as manipulações ocorrem, as consultas vão sendo executadas. A Figura 43 mostra essas duas opções na conexão.

Figura 43: Configurações de modo de conexão do SQL com Power BI

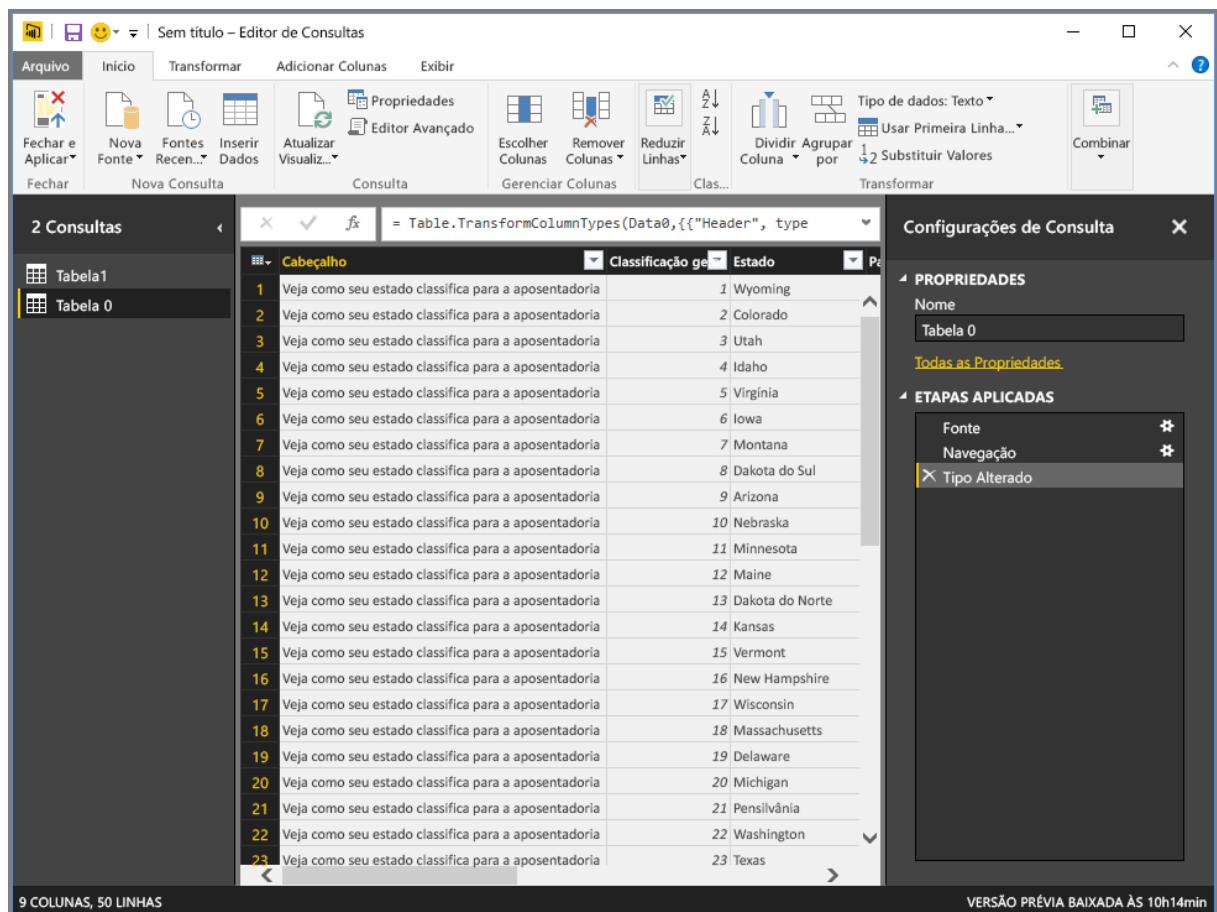


Fonte: Microsoft, 2019.

4.3.1.2 Manipulação e Tratamento de Dados – Nível de Dados e Editor de Consultas

O Editor de Consultas do Power BI é o principal recurso para tratamento e manipulação de dados. Ele permite formatar e combinar dados, agrupar linhas, criar colunas personalizadas, dinamizar colunas, aplicar fórmulas de consulta. Por meio de Etapas, as transformações são aplicadas, e ficam na guia do lado direito. Existem funcionalidades automáticas disponíveis na barra superior da ferramenta. A Figura 44 a seguir mostra a tela contendo essas funcionalidades.

Figura 44: Editor de Consultas do Power BI.



Fonte: Microsoft, 2019.

Com exceção das tarefas agendadas no SQL Agent e a lógica implementada no PLC, todo o tratamento de dados da ferramenta de monitoramento dos dados foi realizado no Power BI, muito concentrada no Editor de Consultas. Um recurso importante é o Editor Avançado, o qual possibilita, por meio da Linguagem de

Fórmulas do Power Query, aplicar expressões personalizadas nas etapas de edição. A Figura a seguir mostra esse recurso.

Figura 45: Editor Avançado do Power BI.



Fonte: Microsoft, 2019.

Na aba de dados é possível visualizar as tabelas contendo os dados, obtidos das fontes conectadas, no caso, via importação. Nessa guia, é possível fazer algumas manipulações tais como renomear colunas, nome de tabelas, inserir novas colunas, classificar dados e filtrar, sendo que estas duas últimas operações são somente a nível de visualização, não interferindo nos dados enviados para o modelo.

Figura 46: Guia de dados do Power BI.

TABELA: Area01 (2.097.146 linhas)

Fonte: Autor, 2019.

A guia de relatório, juntamente com a de dados, permite criar medidas, via fórmulas DAX (Data Analysis Expressions). Trata-se de um conjunto de funções, operadores e constantes que podem ser utilizados em fórmulas, extraindo informações a partir de dados. A Figura 47 abaixo ilustra uma medida criada utilizando DAX.

Figura 47: Exemplo de uma medida criada utilizando DAX.

```
1 Perc_nok_pressao = COUNTROWS(FILTER(Area01;Area01[STATUS_PRESSAO]="NOK"))/COUNTROWS(Area01)*100
```

Fonte: Autor, 2019.

4.3.1.3 Criação de relacionamentos – Nível Modelo

Na aba de Modelo do Power BI, é possível visualizar todas as tabelas inseridas via conexão com fonte de dados, ou criadas no próprio ambiente do Power BI. Esse recurso permite a criação de relacionamentos entre tabelas, por meio de identificadores chave, comuns às duas tabelas. Outra opção é ativar o modo de detecção automática de relações.

A cardinalidade das relações pode ser: “Muitos pra Um” (uma tabela possuindo uma instância de cada valor, enquanto a outra possui várias instâncias do mesmo

valor), “Um pra um” (apenas uma instância de cada valor nas duas tabelas relacionadas) ou “Muitos para Muitos” (várias instâncias de cada valor em ambas as tabelas).

A direção do filtro cruzado nas relações pode ser “Única”, o que quer dizer que a filtragem funciona na tabela em que os valores são agregados, ou “Ambas”, a qual as tabelas são tratadas como se fossem uma única.

Neste projeto, esse item é extremamente importante para garantir a rastreabilidade dos lotes do produto em cada etapa do processo. A Figura 48 mostra um exemplo de relacionamento entre as tabelas Área 01 e Área 02.

Figura 48: Relacionamento entre tabelas na guia de Modelo do Power BI.



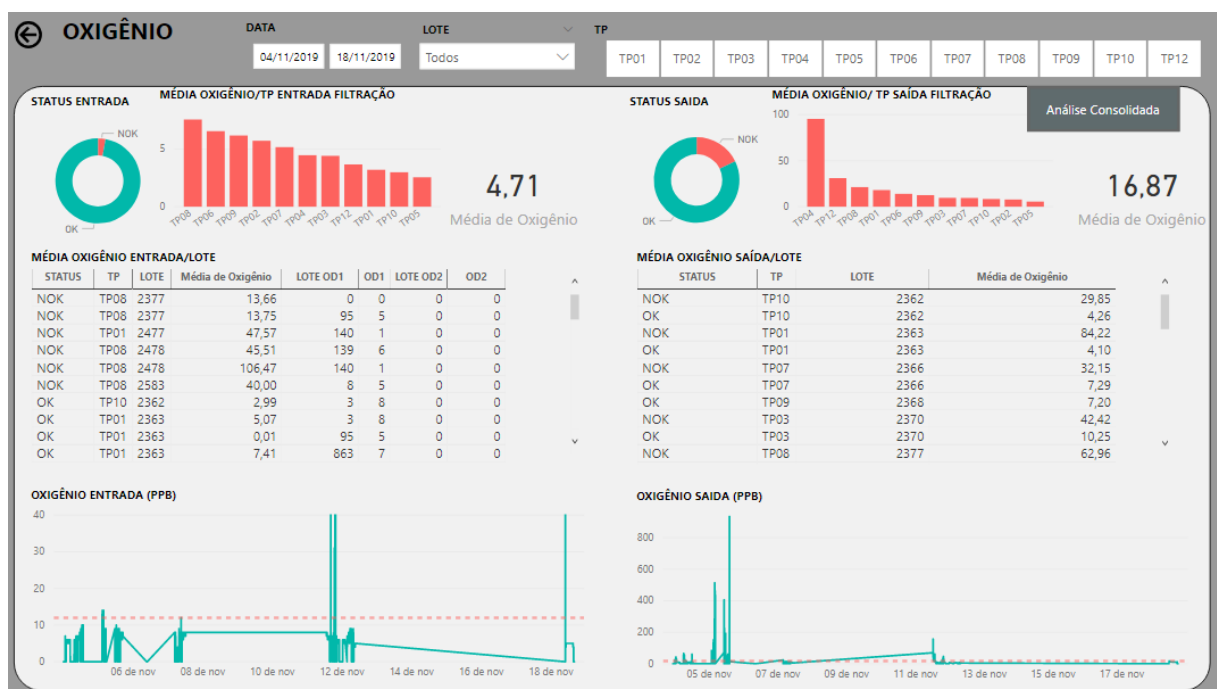
Fonte: Autor, 2019.

4.3.1.4 Criação de Visuais – Nível Relatório

Com os dados inseridos e tratados, é possível criar visuais de maneira muito fácil e rápida, na página de relatório. Os recursos são muitos: gráficos de barra, de dispersão, indicadores, visualização de KPIs, curvas, cartões com exibição de média,

mediana, desvio padrão, botões de navegação entre relatórios, etc. Aqui, a grande vantagem é a possibilidade de se fazer quase tudo que é preciso em um bom dashboard de maneira automática, com um design limpo e claro. Além disso, é possível criar seus próprios modelos de visuais, ou importar do Marketplace, um espaço de compras de produtos e soluções do Power BI, da própria comunidade de usuários. A Figura 49 a seguir mostra um relatório criado no projeto, contendo gráficos de percentual, curvas com seus respectivos valores de setpoint, máximo e mínimo, filtros de data, lote e número de identificação do tanque do processo. Ao lado direito, na guia Visualizações, estão todos os visuais padrões do Power BI, descritos anteriormente, que podem ser utilizados e customizados. Já a Figura 42 apresenta um modelo de relatório de análises criado na ferramenta, com o objetivo de fornecer indicadores de desempenho de processo.

Figura 49: Exemplo de relatório criado no Power BI.



Fonte: Autor, 2019.

Figura 50: Exemplo de análises criadas no Power BI.



Fonte: Autor, 2019.

4.4 Predições de Desvios

Depois de consolidar todos os bancos de dados, abrangendo um grande volume de informações, iniciou-se a etapa de estudos de modelos de predições que agregassem valor para a tomada de decisão da empresa em questão. Para tal, foram utilizados os conceitos de modelos preditivos de Machine Learning. Por se tratar de uma tarefa complexa, decidiu-se por iniciar um estudo de modelos de predição por processos micro, testando e validando a efetividade dos resultados e os possíveis retornos benéficos para a empresa. Desta maneira, as soluções apresentadas nesse trabalho compõem-se de um protótipo inicial para implementação efetiva do Aprendizado de Máquinas nos processos da indústria.

Conforme foi visto na revisão de bibliografia deste trabalho, um algoritmo de Machine Learning é treinado com base em dados para realização de predições, descrições e agrupamentos. Desta maneira, convencionou-se utilizar os mesmos dados coletados para a construção dos painéis e dashboards da ferramenta de BI já descrita anteriormente. Como ação inicial, foi necessário estudar quais informações mais relevantes de aprendizado seriam possíveis e como aplicá-las com os dados

disponibilizados. Optou-se pela criação de algoritmos capazes de realizar a predição de indicadores de qualidade do produto acabado, de modo a atuar no processo antes da efetivação do problema. No cenário atual, a empresa não possui nenhum algoritmo implementado voltado para tal fim e existem dificuldades em relacionar os desvios no produto acabado com os itens de controle de processo. Desta maneira, a primeira classificação do projeto relaciona-se com este item de qualidade, de modo que os atributos serão as variáveis de processo, tais como o percentual dentro de faixa de temperatura, pressão e incorporação de oxigênio, enquanto que as classes serão os desvios de qualidade apontados no produto final.

Embora o objetivo final seja a aplicação das predições para tal fim, convencionou-se inicialmente a aplicação dos algoritmos em bases de dados já consolidadas e disponíveis na Web, para fins acadêmicos e validação de métodos passíveis de serem utilizados no caso real. Com o estudo e entendimento destas bases, realiza-se então a aplicação para os dados de processo coletados neste trabalho, comparando os resultados obtidos entre eles. Assim, a base escolhida denomina-se “Red Wine”¹ e está disponível na página da UCI, um dos mais populares repositórios de bases de dados para Machine Learning.

As etapas do desenvolvimento são descritas a seguir.

4.4.1.1 Linguagem Python e pacote Pandas

Para o desenvolvimento dos algoritmos apresentados nesse trabalho foi utilizada a linguagem Python. Trata-se da linguagem mais utilizada na área de Ciência de Dados, uma vez que possui uma grande comunidade ativa, com inúmeras colaborações na Web que fortalecem seu uso. Além disso, permite a criação de visuais de maneira simples e fácil, o que é importante para a entrega da informação coerente da melhor maneira para os clientes do projeto, importante para a efetividade nas tomadas de decisão. Outra razão é a grande quantidade de bibliotecas existentes, que permitem soluções integradas.

¹ Disponível em < <https://archive.ics.uci.edu> > Acesso em: Dez. 2019

Outra peça fundamental é o pacote Pandas, que oferece todo suporte para análises e manipulação de dados. De maneira simplificada, a biblioteca Pandas permite a manipulação de DataFrames, vetores e matrizes, essenciais para processamento dos dados. Além disso, possibilita o pré-processamento de dados, com soluções para preenchimento de dados faltantes e nulos, transformação de vetores, mantendo o arquivo original dos dados intacto. Outras funcionalidades importantes são juntar e concatenar vetores, aplicação de filtros, cálculos de média, desvio padrão e variação automáticos, aplicação de índice automático, etc. Sua instalação é realizada utilizando o próprio mecanismo de download de pacotes do Python.

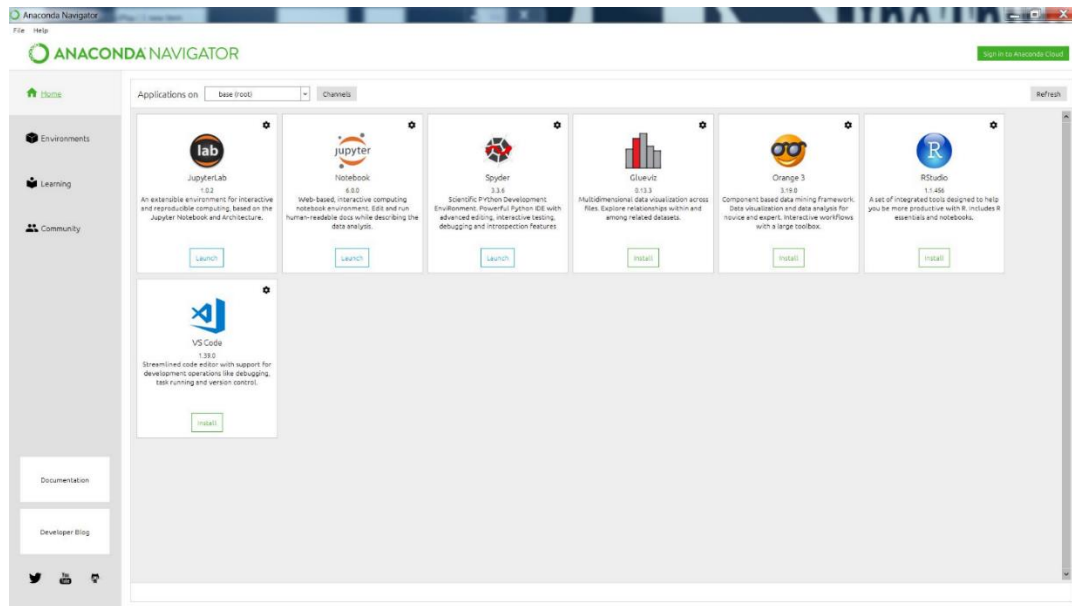
4.4.1.2 Scikit-Learn

Com a escolha da linguagem em Python, optou-se também pela utilização da biblioteca Scikit-Learn, com base fundamental para realização dos algoritmos de aprendizado. Essa biblioteca utiliza o pacote Pandas, e possui suporte para algoritmos de classificação, regressão e agrupamento tais como árvores de decisão, máquinas de vetores de suporte, florestas aleatórias, etc. Trata-se de uma ferramenta de uso livre, inclusive para usos comerciais, sendo esta também uma grande vantagem para utilização neste projeto

4.4.1.3 Anaconda Navigator

Anaconda Navigator é uma interface gráfica para desktop com diversas soluções em Python integradas. Sua grande vantagem é a abrangência de inúmeros pacotes usuais em programação de algoritmos de aprendizado de máquinas, de modo que não é necessário preocupar-se com as versões e atualizações destes pacotes. Além disso, possibilita um ambiente mais visual para programação, fora do terminal de comandos básico do sistema operacional.

Figura 51: Ambiente do Anaconda Navigator.



Fonte: Anaconda, 2019.

Em conjunto com o Anaconda, foram utilizadas as ferramentas Spyder e Jupyter Notebook. O Spyder é um ambiente de desenvolvimento e debug de códigos em Python, exploração de dados e visualizações de resultados em gráficos e tabelas. Já o Jupyter oferece as mesmas funcionalidades e ainda permite o desenvolvimento em ambiente Web, facilitando o compartilhamento e trabalho colaborativo.

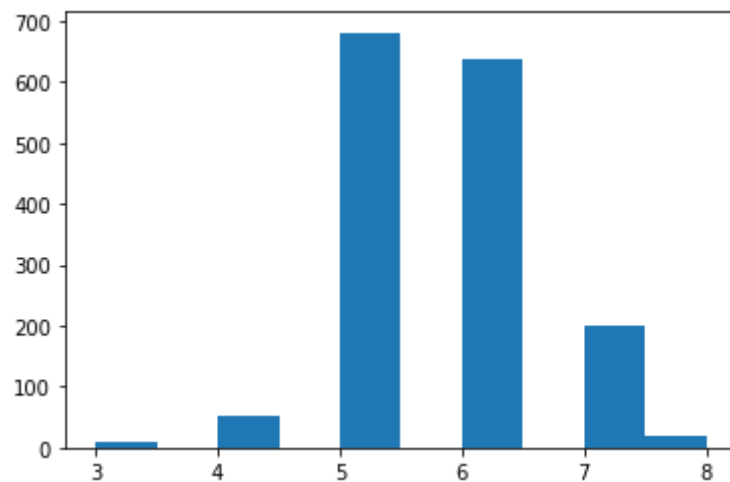
4.4.1.4 Aplicação de Modelos de Classificação Preditiva

Conforme apresentado, os algoritmos de classificação fazem parte dos modelos preditivos, e são caracterizados por conterem entradas com rótulos atribuídos à elas. Utilizou-se um modelo inicial de estudo da qualidade de um produto, por meio de um banco de dados externo, e outro conjunto de dados pertinente aos dados coletados neste projeto, para entendimento de desvios no produto acabado, de modo que um determinado atributo foi avaliado em relação aos parâmetros de processo.

A utilização de dados não referentes à empresa é justificada como forma de um estudo inicial e comparativo dos resultados. Este conjunto de dados é constituído de

informações referentes a um produto e sua classe é definida pelo rótulo Qualidade, o qual possui um valor inteiro entre 3 e 8, conforme mostra o histograma da Figura 52.

Figura 52: Histograma dos dados de Qualidade.



Fonte: Autor, 2019.

Os parâmetros são descritos em outras 11 colunas, conforme mostra a descrição geral dos dados na Figura.

Figura 53: Visão geral dos parâmetros de um dos conjuntos de dados.

```

fixed acidity volatile acidity citric acid ... sulphates alcohol quality
0          7.4           0.70      0.00 ...      0.56      9.4        5
1          7.8           0.88      0.00 ...      0.68      9.8        5
2          7.8           0.76      0.04 ...      0.65      9.8        5
3         11.2           0.28      0.56 ...      0.58      9.8        6
4          7.4           0.70      0.00 ...      0.56      9.4        5

[5 rows x 12 columns]
```

Fonte: Autor, 2019.

Para os dados do processo industrial, a classe escolhida infere-se a existência ou não de determinado desvio no produto, com o objetivo de prever a possibilidade de tal característica antes da finalização do processo. Foram utilizados três parâmetros iniciais, sendo estes o percentual dentro de faixa aceitável para a temperatura, contrapressão e oxigênio incorporado em determinada fase de fabricação. A Figura 54 mostra parte da tabela contendo os dados na respectiva ordem.

Figura 54: Parâmetros do conjunto dos dados de processo.

	0	1	2
0	68.6659	92.1111	100
1	100	99.6358	30.2206
2	89.1046	84.389	91.4649
3	100	100	1.71479
4	85.4284	100	90.8317
5	99.8062	100	53.13
6	100	99.4936	54.7303
7	88.386	99.582	41.7348
8	99.9808	99.9936	42.5239
9	51.0026	59.5466	2.62267
10	100	100	2.41901

Fonte: Autor, 2019.

A primeira etapa é constituída de realizar o pré-processamento dos dados. Para isso, foi realizado:

- Tratamento de dados nulos e zerados: para os dados sem identificação de lote foi realizado a exclusão do dataset, enquanto que variáveis de processo com valores nulos foram preenchidas com a média da batelada em questão;
- Tratamento de outliers: para os valores de outliers foi aplicado um estimador de dados da biblioteca sklearn,
- Cálculo do percentual dentro de faixa de cada variável: para consolidação do grande volume de dados referentes a cada lote foi aplicado o percentual como medida de entrada do algoritmo;
- Concatenação de tabelas: foi realizada a associação das variáveis utilizadas de cada etapa produtiva selecionada, utilizando os lotes como identificador único;
- Tratamento das classes atribuídas: os dados do atributo escolhido como classe foram analisados e organizados em uma tabela, utilizando também o lote como identificador único. A classificação foi definida como 0 ou 1, ou seja, 0 o produto não apresentou o desvio e 1 como a presença do mesmo;
- Normalização e padronização de dados: utilizando a biblioteca sklearn os dados foram normalizados e padronizados, fator importante para aumentar a acurácia do algoritmo.

A Figura 55 abaixo mostra o pré-processamento aplicado para os dados do processo

Figura 55: Pré-processamento dos dados de processo.

Nome	Tipo	Tamanho
ads	DataFrame	(72, 2)
base	DataFrame	(1080989, 21)
base_cent	DataFrame	(1403795, 11)
base_ok_oxi_cent	Series	(106,)
base_ok_oxi_cent2	Series	(11,)
base_ok_pressao	Series	(145,)
base_ok_temp	Series	(140,)
classe_teste	int64	(11,)

```

5 @author: Bruna
6
7 import pandas as pd
8 import numpy as np
9 base = pd.read_csv('ferm.csv')
10 base_cent = pd.read_csv('centr.csv')
11 ads = pd.read_csv('ads.csv')
12 base.head()
13 base.describe()
14
15 pd.isnull(base['Temperatura'])
16 base.loc[pd.isnull(base['SP_Pressao'])]
17 base.loc[pd.isnull(base['Temp_max'])]
18 base_cent.loc[pd.isnull(base_cent['Lote_FermCentr2'])]
19
20 base.drop(base[pd.isnull(base['Lote'])].index, inplace=True)
21 base.drop(base[base.Lote == 0].index, inplace=True)
22 base.drop(base[pd.isnull(base['Temperatura'])].index, inplace=True)
23 base.drop(base[pd.isnull(base['Pressao'])].index, inplace=True)
24 base.drop(base[pd.isnull(base['Temp_Cone'])].index, inplace=True)
25 base.drop(base[pd.isnull(base['Produto'])].index, inplace=True)
26 base.drop(base[pd.isnull(base['Tanque'])].index, inplace=True)
27
28 base.Temperatura = base.Temperatura/100
29 base.SP_Temp = base.SP_Temp/100
30 temp_min = base.SP_Temp - 0.5
31 temp_max = base.SP_Temp + 0.5
32 base['Temp_min'] = temp_min
33 base['Temp_max'] = temp_max
34 base.drop(base[pd.isnull(base['SP_Temp'])].index, inplace=True)
35 base.drop(base[pd.isnull(base['Temp_min'])].index, inplace=True)
36 base.drop(base[pd.isnull(base['Temp_max'])].index, inplace=True)
37
38 base.Pressao = base.Pressao/100
39 base.SP_Pressao = base.SP_Pressao/100
40 pressao_min = base.SP_Pressao - 0.02
41 pressao_max = base.SP_Pressao + 0.02
42 base['Pressao_min'] = pressao_min
43 base['Pressao_max'] = pressao_max
44

```

Fonte: Autor, 2019.

Após realizado o pré-processamento, foi aplicado o modelo de classificação K-Nearst Neighbors, com o auxílio da biblioteca do Scikit-Learn. Este modelo foi escolhido por sua simplicidade e baixa complexidade de aplicação, como forma inicial de estudo. A Figura 56 ilustra alguns trechos de códigos aplicado para tal modelo.

Figura 56: Aplicação do método K-NN.

```

36 from sklearn.model_selection import train_test_split
37 X_train, X_test, y_train, y_test = train_test_split(predictors_wine, reviews, test_size=0.15, random_state=1, stratify=reviews)
38
39 from sklearn.neighbors import KNeighborsClassifier
40 from sklearn.model_selection import GridSearchCV
41 knn2 = KNeighborsClassifier()
42 param_grid = {'n_neighbors': np.arange(1, 25)}
43 knn_gscv = GridSearchCV(knn2, param_grid, cv=3)
44 knn_gscv.fit(predictors_wine, reviews)
45 knn_gscv.best_params_
46
47 knn = KNeighborsClassifier(n_neighbors=13, metric='minkowski', p=2)
48 knn.fit(X_train, y_train)
49 previsoes = knn.predict(X_test)
50
51 from sklearn.metrics import confusion_matrix, accuracy_score, recall_score, precision_score
52 precisao = accuracy_score(y_test, previsoes)
53 matriz = confusion_matrix(y_test, previsoes)
54 recall2 = recall_score(y_test, previsoes, pos_label='2')
55 precision = precision_score(y_test, previsoes, pos_label='2')
56

```

Fonte: Autor, 2019.

4.5 Metodologia de Validação

Conforme já exposto no início deste capítulo, o trabalho aqui apresentado constitui-se de um projeto ainda em desenvolvimento, de modo que os resultados apresentados deste projeto não se encontram ainda totalmente integrados. Desta maneira, a metodologia aqui adotada será a validação de cada parte constituinte do projeto. A seguir, detalha-se os pontos avaliados para tal.

- Validação da Coleta e Banco de Dados: para verificar se a coleta está sendo realizada da maneira correta, será feita uma comparação dos valores disponibilizados no RunTime do IGS com aqueles disponibilizados nas tabelas de visualização online do Step7, seguida de uma solicitação de leitura de tais dados salvos no Banco de Dados;
- Validação dos Dashboards no Power BI: será detalhado o passo a passo da construção de um painel e um caso de sucesso do uso da ferramenta;
- Validação das Predições: serão apresentados dois cenários de aplicação de modelos de predição, utilizando acurácia, precisão e sensibilidade como meios de validação.

5 RESULTADOS

A criação da ferramenta de monitoramento e análise de indicadores foi de maneira contínua e integrada com as equipes da área de aplicação do projeto. Um brainstorming guiado com a equipe de supervisão foi realizado no início do projeto, para captação das ideias e entendimento de algumas partes técnicas do processo de produção. Essa atividade permitiu que os primeiros indicadores fossem idealizados e quais dados seriam necessários para seu desenvolvimento.

Toda a etapa de desenvolvimento foi realizada em conjunto com as equipes de automação e TI, com a entrega dos dashboards e relatórios iniciais sendo avaliada e criticada pela equipe de supervisão e coordenação, gerando melhorias e criando padrões próprios.

5.1 Validação da Coleta e Banco de Dados

A estrutura para o fluxo de dados, desde sua coleta nos PLC's, passando pelo servidor e cliente OPC e sendo armazenado no Banco de Dados SQL Server mostrou-se efetiva e sem nenhum tipo de erros ou falhas capazes de prejudicar a capacidade e o objetivo final do sistema proposto no trabalho. Durante a implementação do projeto, foram realizadas constantes comparações dos valores disponibilizados nos ambientes do Step7, do IGS, OPC Data Logger com o valor armazenado no Banco de Dados, de modo que a integridade do dado a ser armazenado foi mantida.

5.2 Estudo de caso do passo a passo de desenvolvimento e resultados obtidos com o uso da ferramenta

Um dos relatórios criados no dashboard desenvolvido nesse projeto é a análise de variáveis da etapa de fermentação, fase muito importante para a qualidade do produto final. No momento inicial de desenvolvimento desse painel, foi realizada reunião de alinhamento de expectativas dos supervisores e gerentes responsáveis por essa etapa produtiva. Foi identificado a necessidade de monitoramento e análise

de quatro variáveis de processo: a temperatura e contrapressão do tanque, a temperatura no cone e o extrato instantâneo.

Com a escolha das variáveis em mãos, a equipe de automação responsabilizou-se por realizar toda a configuração citada no capítulo de desenvolvimento desse trabalho. Para isso, identificou-se que os dados de temperatura seriam possíveis com a extração da leitura dos sensores de temperatura localizados no interior e no cone dos tanques. Para o extrato, o valor deveria ser calculado a partir de variáveis de entrada, tais como massa e volume do tanque. Além disso, seria necessário a leitura de alguns bits de processo para identificar as etapas em que os dados seriam enviados para o sistema. Todos os valores de setpoint também foram considerados importantes de serem coletados do Step7, diretamente das receitas.

Definidas as variáveis, a etapa subsequente foi o tratamento de alguns dados no ambiente do Step 7. Deste modo, foi criada uma lógica para enviar os dados para uma instância de função específica do projeto somente a partir do início de enchimento do tanque, já contendo um lote específico para este processo, além de um contador de tempo de fermentação.

Posteriormente, a coleta de dados via OPC foi realizada, utilizando o Server (IGS) com tempo de coleta de 500ms, enquanto que no Client (OPC Datag Logger) foi utilizado um tempo de 1s, visando garantir uma boa sincronia de envio e recebimento de dados.

A próxima ação foi a criação das tabelas no Banco de Dados operacional. Optou-se por criar uma única tabela para este estágio do processo, alimentada sequencialmente por todos os dados dos tanques em operação. A identificação é realizada pelos campos de lote e tanque, além da data e hora de coleta dos dados. No SQL Server, utilizando o SQL Agent, foram criadas rotinas de limpeza e tratamento de dados, tais como excluir outliers, nulos e zeros.

Finalmente, com a conexão configurada com o Banco de Dados, iniciou-se a fase de modelagem e criação das análises dos dados, no Power BI. Para o relatório descrito nesse exemplo foram utilizadas curvas, contendo a faixa de operação aceitável e as análises de percentual em relação à esses parâmetros. A Figura 57 mostra o relatório criado.

Figura 57: Relatório criado para análise de curvas no Power BI

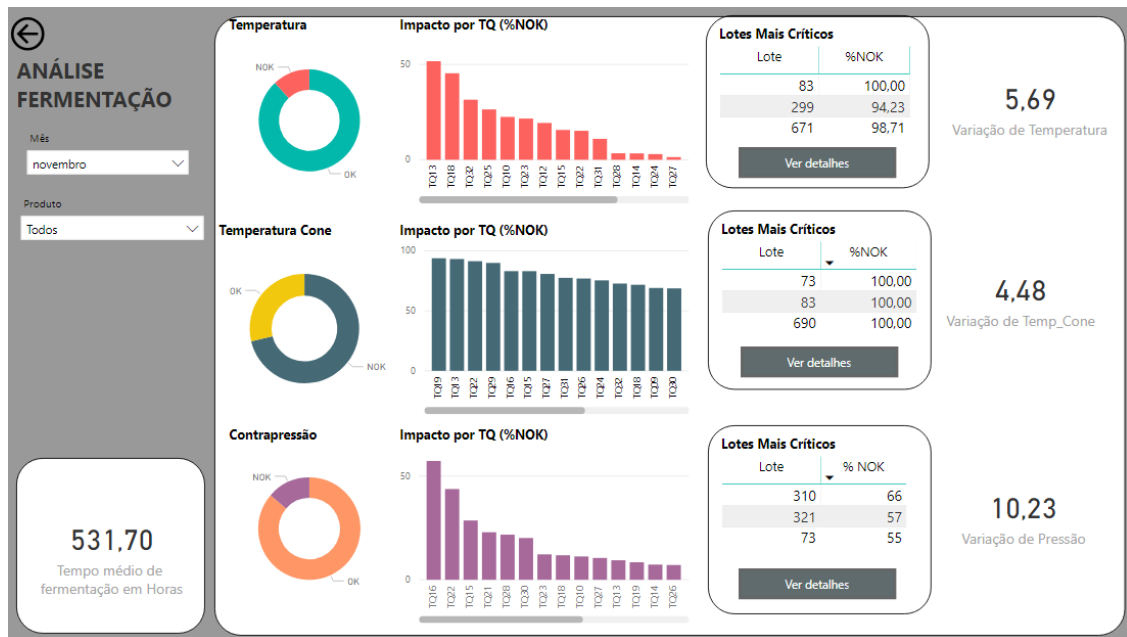


Fonte: Autor, 2019.

Nas curvas do painel acima, a cor preta representa o valor do setpoint da variável, enquanto que a vermelha apresenta os valores máximo e mínimo aceitáveis para a mesma, e a cor verde representa o valor real. Na curva de temperatura, por exemplo, é possível observar uma mudança de setpoint de 10°C para 15°C conforme receita de fabricação do produto. Desta maneira, o controlador reduz a abertura da válvula de entrada de etanol para a camisa do tanque, diminuindo o nível de resfriamento do tanque. Como não há aquecimento, a temperatura realiza comportamento de rampa, incrementando gradualmente. Para o caso da contrapressão, observa-se oscilações devido à relação entre aumento da pressão no tanque com os momentos de alívio do mesmo.

A empresa em questão utiliza as metodologias PDCA (Plan, Do, Check, Act) e 5 Porquês para resolução de problemas, de modo que o sistema desenvolvido deveria ser integrado com essas ferramentas. Desta maneira, realizou-se a construção de análises baseadas em Paretos, para identificação dos maiores impactos. A Figura 58 mostra um exemplo dessa visualização, em que os indicadores são generalizados por tipo de produto na escala mensal, apresentando os tanques com maior percentual de parâmetros fora de faixa, assim como os lotes mais críticos.

Figura 58: Análises no Power BI.



Fonte: Autor, 2019.

Com a utilização da ferramenta, os supervisores puderam analisar o comportamento do processo, visualizando o percentual dentro de faixa e a evolução dos indicadores ao longo da produção. Com essas informações, eles conseguem visualizar se os problemas existentes são pontuais ou crônicos, se está concentrado em uma etapa ou generalizado, se um tanque apresenta sempre os mesmos erros nos processos, se existe a necessidade de alteração de uma receita ou não, etc.

Um caso de sucesso de resolução de problemas no nível operacional com o uso da ferramenta foi a identificação de problemas em válvulas. O operador verificou que a temperatura do tanque estava caindo drasticamente, não conseguindo mais atingir o valor de setpoint demandado pelo controlador. Na tela de controle do sistema supervisorio (já existente em área), o operador verificou que todos os parâmetros de processo estavam corretos, a produção estava no passo certo da receita e não havia nenhum indício de problema de controle. Desta maneira, verificando em campo, constatou-se que as válvulas de entrada de etanol na camisa do tanque, responsáveis pelo resfriamento do mesmo, estavam danificadas, permitindo a passagem de maiores volumes de líquido. Como bloqueio de efeito da anomalia, a válvula foi reparada pela equipe de manutenção. Para que o problema não voltasse a ocorrer era necessária uma análise que justificasse o desgaste da peça, ou seja, se tratava-

se de deterioração natural (pelo tempo de utilização) ou forçada (utilização e manutenção incorretas). Nessa avaliação, constatou-se que uma deterioração forçada, causada pela abrupta abertura e fechamento da válvula ao longo do tempo. Uma maneira de reduzir o desgaste causado por essa reação é a instalação de reguladores de fluxo na entrada da válvula, ação realizada no tanque em que foi constatada a anomalia e em todos os demais da área produtiva, executando as devidas manutenções corretivas nas demais. Para o fechamento do ciclo da resolução do problema, foi criado um plano de manutenção contemplando essas práticas.

Da observação deste caso, nota-se que o sistema desenvolvido e apresentado nesse trabalho é uma importante ferramenta para identificação de problemas e, com o auxílio de modelos como o do 5 Porquês e a metodologia PDCA.

5.3 Resultados dos modelos de predição

Conforme discutido anteriormente, foram utilizados dois conjuntos de dados, sendo um destes um banco de dados disponível na Web, e outro proveniente dos dados do processo industrial analisado neste trabalho. Estudos já desenvolvidos deste trabalho mostram uma acurácia máxima de aproximadamente 76% quanto ao modelo de classificação, valor próximo ao obtido neste trabalho. Entretanto, um indicador importante seria a Sensibilidade, uma vez que mede o percentual de observações positivas classificadas corretamente. A tabela a seguir mostra o comparativo entre os resultados obtidos entre os dois conjuntos de dados, enquanto que as figuras mostram a matriz de confusão de cada um deles.

Tabela 7: Resultados dos testes de predição.

Conjunto de Dados	Acurácia (%)	Precisão (%)	Sensibilidade (%)
Dados de qualidade do produto (didático)	76,3	75,9	81,25
Dados de processo (real)	81,8	100	55,5

Fonte: Autor, 2019.

Figura 59: Matriz de Confusão dos Dados de Qualidade (didático).

	0	1
0	79	33
1	24	104

Fonte: Autor, 2019.

Figura 60: Matriz de Confusão dos Dados do processo real.

	0	1
0	13	0
1	4	5

Fonte: Autor, 2019.

A diferença das medidas obtidas é resultante da quantidade de parâmetros e de instâncias disponíveis. Para o caso dos dados de processo, ainda não se tem um grande volume disponível para validação da eficiência deste modelo. Analisando os resultados destes conjuntos de dados, observa-se que tal método de classificação possivelmente não é o mais indicado para tal caso, sendo necessário uma opção que contemple a complexidade do problema. Ressalta-se, porém, a necessidade de contemplar mais parâmetros para obtenção de um modelo mais eficiente.

6 CONCLUSÃO

Este trabalho propôs-se à criação de uma ferramenta de monitoramento de informações de controle de processo de uma área industrial, para fornecer análises complexas de dados referentes às variáveis de produção auxiliando na tomada de decisão da organização. Com a utilização da ferramenta, obteve-se resultados de sucesso na identificação de problemas, a nível operacional, e nas tomadas de decisão, a nível de supervisão e gerência. No cenário anterior à existência dessa ferramenta, os operadores não tinham acesso à um histórico detalhado dos dados de processo, com a visualização de gráficos de baixa qualidade da informação nas telas de supervisorio. Nestes sistemas, a enorme quantidade de dados não permite velocidade de identificação e atuação em problemas de campo.

No nível tático, supervisores e gerentes passaram a ter uma visão comparativa entre curvas de processo, identificando anomalias recorrentes em todos os lotes, informações que não eram possíveis de serem visualizadas com velocidade apenas com os sistemas existentes. Mudanças em receitas e no tempo de utilização de insumos foram tomadas visando a melhoria na qualidade do produto final.

Com relação à parte de predição deste trabalho, segue-se em desenvolvimento e estudo e, por tratar-se de uma solução mais complexa, demanda-se mais aplicações de modelos, seguidas de constantes testes para a total integração do projeto proposto. No entanto, sua completa finalização apresentará inúmeras vantagens para o ambiente empresarial proposto.

7 REFERÊNCIAS BIBLIOGRÁFICAS

ANDREONI, Martin. **An Intrusion Detection And Prevention Architecture For Software Defined Networking**, 2014

ANTONELLI, Ricardo Adriano. **Conhecendo o Business Intelligence. Uma Ferramenta De Auxílio à Tomada De Decisão** – Revista TECAP – Número 03, Ano 03, Volume 03, 2009.

BARBIERI, CARLOS. **BI2 - Business Intelligence – Modelagem e Qualidade**. Elsevier, 2011.

BONEL, Claudio. **Afinal o que é Business Intelligence: Do conceito a aplicação**. Rio de Janeiro: Perse, 2017.

BOTELHO, Fernando R.; FILHO, Edelvino Razzolini. **Conceituando o termo Business Intelligence: Origem e principais objetivos**. Universidade Federal do Paraná, 2014.

CHEN, H., CHIANG, R. H. L., STOREY, V. C. **Business intelligence and analytics: From big data to big impact**. MIS Quarterly, 2012.

CZERNICKI, B. **Next Generation Business Intelligence Software with Silverlight 3**. New York: Apress, 2009.

ELENA, C. **Business Intelligence**. Journal of Knowledge Management, Economics and Information Technology, 2011.

ELIAS, Diego. **O que significa OLTP e OLAP na prática**. Disponível em: <<https://canaltech.com.br/business-intelligence/o-que-significa-oltp-e-olap-na-pratica>> Acesso em: 01 de dez. 2019.

GÉRON, A. (2017). **Hands-On Machine Learning with Scikit-Learn and TensorFlow**. United States of America: O'Reilly Media, Inc., 1005 Gravenstein Highway North, Sebastopol.

GOODFELLOW, I., BANGIO, Y. & COURVILLE, A. **Deep Learning: MIT Press**. 2016

GUEDES, Anselmo. **BI Operacional: o BI de próxima geração – Parte 01**. 2013. Disponível em: <<https://imasters.com.br/devsecops/bi-operacional-o-bi-de-proxima-geracao-parte-01>> Acesso em: 28 de nov. 2019.

GUIDO, S., & MULLER, A. (2016). **Introduction to Machine Learning with Python**. United States of America.: O'Reilly Media, Inc., 1005 Gravenstein Highway North, Sebastopol.

LUHN, H. P. **A Business Intelligence System**. IBM Journal of Research and Development, 1958.

MATOS, David. **Power BI ou Tableau?**. 2019. Disponível em: <www.cienciaedados.com/power-bi-ou-tableau> Acesso em: 03 de dez. 2019.

NALANI, Avnashi. **NKK Classification using Scikit-learn**. 2018. Disponível em: <<https://www.datacamp.com/community/tutorials/k-nearest-neighbor-classification-scikit-learn>> Acesso em: 12 de nov. 2019.

PRIMAK, Fábio V. **Decisões com B.I. (Business Intelligence)**. Rio de Janeiro: Ciência Moderna, 2008

RIBEIRO, Rogério F. **Business Intelligence como garantia de diferencial competitivo**. UNIMINAS, 2005.

SILVA, C. Josenildo. **Algoritmo de Aprendizado de Máquinas: qual deles escolher?**. 2018. Disponível em: <<https://medium.com/machina-sapiens/algoritmos-de-aprendizagem-de-m%C3%A1quina-qual-deles-escolher-67040ad68737>> Acesso em: 27 de nov. 2019.

SOUZA, Cesar. **Análise de Poder Discriminativo através de Curvas ROC**. 2009. Disponível em: < <http://crsouza.com/2009/07/13/analise-de-poder-discriminativo-atraves-de-curvas-roc> > Acesso em: 05 de dez. 2019.

SRIVASTAVA, Tavish. **Introduction to Online Machine Learning: Simplified**. 2015. Disponível em: <<https://www.analyticsvidhya.com/blog/2015/01/introduction-online-machine-learning-simplified-2/>> Acesso em: 03 de dez. 2019.

STEFFINE, Gregory P. **Hyper: Changing the way you think about, plan and execute business intelligence for real results, real fast!**. Sanderson Press, 2015.

TAN, Pang-Ning, STEINBACH, Michael, KARPATNE, Anuj & KUMAR, Vipin. **Introduction to Data Mining**. 2. ed. Michigan, USA: Pearson, 2005.

TURBAN, E. **Business intelligence: um enfoque gerencial para a inteligência do negócio**. São Paulo: Bookman, 2009.

VERCELLIS, C. **Business Intelligence: Data Mining and Optimization for Decision Making** (p. 417). United Kingdom: Wiley, 2009.