DOI: 10.48011/asba.v2i1.1147

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

Bruna Corrêa de Morais, Renato Ferreira Fernandes Jr, Renato Santos Carrijo

Universidade Federal de Uberlândia, Uberlândia, Brasil
E-mails: morais.brunacorrea@gmail.com, rffernandes@ufu.br, rscarrijo@ufu.br

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.

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 em dados históricos. O sistema desenvolvido utilizou-se dos conceitos de Inteligência de Negócios para a construção de telas interativas, 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.

Keywords: Business Intelligence, Analysis Industrial Process, Machine Learning, Prediction

Palavras-chaves*:* Inteligência Negócios, Análise Processo Industrial, Aprendizado de Máquinas, Predição

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 a inteligência de negócios, o qual vem assumindo importante papel para transformação da eficiência na tomada de decisão nas organizações (STEFFINE, 2015).

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 (TURBAN, 2009).

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 (STEFFINE, 2015).

Este 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. INTELIGÊNCIA DE NEGÓCIOS E APRENDIZADO DE MÁQUINA

As tecnologias de inteligência de negócios estão cada vez mais sendo utilizadas dentro das empresas para melhorar a qualidade e eficiência de seus processos. Desta forma, esta seção fornece uma breve explanação sobre inteligência de negócios e do aprendizado de máquina que são utilizados neste projeto.

Inteligência de Negócios

A Inteligência de Negócios ou *Business Intelligence* (BI), engloba um conjunto de técnicas, conceitos e métodos voltados para apoiar a tomada de decisão de uma organização. O BI utiliza modelos matemáticos e metodologias de análise que utiliza dados para tomadas de decisão complexas (VERCELLIS, 2009).

Dentro da arquitetura de BI é importante o conceito de ciclo da informação que relaciona dados, informação e conhecimento. O ciclo da Informação se inicia com a aquisição de dados, passa para a etapa da geração da informação que responsável pelo agrupamento dos dados, excluindo aqueles que não são relevantes. A próxima etapa é prover conhecimento através da construção de indicadores e métricas capazes de fornecer análises para os tomadores de decisão. Por

fim são tomadas decisões e propostas ações para que iniciativas tenham efeito na organização. Também deve ser acompanhado os resultados para mensurar o desempenho da ferramenta, rever os processos, encontrar possíveis erros que possam estar reduzindo a qualidade da informação (BONEL, 2017).

Na arquitetura típica de BI é necessária uma grande armazenagem de dados através de um armazém de dados (DW *Data Warehouse*) ou um repositório de dados (DM *Data Marts*) (TURBAN, 2009). Em relação a análise dos dados, as metodologias de *Online Analytical Processing* (OLAP) e *Online Transaction Processing* (OLTP) são utilizadas em BI. O OLTP é voltado para a organização e categorização de informações, de modo que possui alta velocidade de manipulação de dados. Por outro lado, o OLAP está voltado para analise (*insights)*, através de uma visualização multidimensional, o que permite uma análise eficaz dos dados sob diferentes perspectivas e com diferentes granularidades (CHEN et.al., 2008).

A respeito da manipulação de dados, de acordo com Sahay e Ranjan (2008) outro processo fundamental em BI são a Extração, Transformação e Carga (ETL). O ETL baseia-se na Identificação da origem dos dados, Limpezas, ajustes, correções de imperfeições, padronização e correção de variabilidades, carregar dados para o DW, atualização do DW. O ETL proporciona dados de qualidade para o DW e DM. Também, o processo de Data Mining visa a identificação de padrões nos dados operacionais, ou derivados de um DW ou DM através do uso de técnicas de inteligência artificial e de aprendizagem estatísticas ou matemáticas para extrair informações uteis dos dados (TURBAN, 2009).

Atualmente existe uma crescente de tecnologias de BI, chamadas BI 2.0, onde tem permitido o avanço do conceito das tecnologias da Indústria 4.0. Existem hoje no mercado diversas soluções de tecnologia de BI que diferem devido a velocidade de processamento de dados, estabilidade de performance, facilidade de uso e custo-benefício. Destacam-se o Power BI da (MICROSOFT, 2019) e o Tableau (SZEWRAŃSKI, et al., 2017).

Aprendizado de Máquina

O Aprendizado de Máquina ou *Machine Learning (ML)* é o estudo de métodos computacionais para automatização de processos. O ML alcançou crescimento significativo nas últimas décadas, com a possibilidade oferecida pelas novas tecnologias, pela internet e pela melhoria contínua na capacidade de processamento das máquinas. 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 (BOSE & MAHAPATRA, 2001).

Segundo Géron (2017), a aplicação do ML dentro do BI traz vantagens como: 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 e redução de custos uma vez que fornece informações reduzidas de erros, evitando desperdícios e melhorando desempenhos.

Em relação aos modelos de ML, eles podem ser classificados de acordo com as técnicas de treinamento em: Aprendizado Supervisionado, Não Supervisionado, Semi supervisionado ou Aprendizado por Reforço. No aprendizado supervisionado, o conjunto de dados de treinamento inclui a solução desejada. No aprendizado não supervisionado, não há uma inclusão de rótulo no conjunto de dados de treinamento. Deste modo, o computador precisará identificar perfis nos dados, encontrando uma representação mais informativa. Neste aprendizado se destacam os algoritmos K-Médias, Clusterização Hierárquica, Máquinas de Bolstzman, t-distributed Stochastic Neighbor Embedding(T-SNE). 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 (GÉRON, 2017).

Em relação a forma de aprendizado, ele pode ser classificado em aprendizado por batelada ou aprendizado online. No aprendizado por batelada se dispõe de uma grande quantidade de dados armazenado previamente e o sistema é treinado de forma offline, uma vez que necessitará de maior capacidade de processamento e tempo. As restrições para o aprendizado por batelada são em sistemas que apresentam mudanças frequentes, com fluxo de dados constante, e para aplicações com fortes restrições de recursos de armazenamento e processamento. No aprendizado online, 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 (BOSE & MAHAPATRA, 2001)

No contexto dos modelos de Aprendizado de Máquinas, existem três principais modelos: Aprendizagem Preditiva, Descritiva ou Híbrida. 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 procedente de um conjunto finito e não ordenado, então é o caso de uma classificação. 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. Por não contemplar o escopo deste trabalho, não será detalhado este modelo (TAN, et.al., 2005).

Os métodos de ML não necessitam de um conhecimento profundo sobre as estrutura do modelo que descreve os dados. Esta característica é útil para modelar o comportamento de sistemas não lineares complexos. Entre os métodos computacionais utilizados, são redes neurais artificias (ANNs), maquinas de suporte vetorial (SVMs), mapas auto-

organizáveis (SOMs) e k-vizinhos mais próximos (K-NN) (VERCELLIS, 2009).

As ANNs buscam simular o processamento de informação realizado pelo cérebro humano através de uma rede de neurônios que passa por uma fase de treinamento dos neuronios. As SVM está relacionada a métodos de aprendizado supervisionado que analisam os dados e reconhecem padrões. O método SOM permite diminuir a dimensão de um grupo de dados, mantendo a representação real de acordo com características relevantes das entradas.

O método *K-Nearst Neighbors* (K-NN) foi um dos primeiros a serem desenvolvidos na área de Aprendizado de Máquinas. Trata-se de um modelo baseado em instâncias, que utiliza do critério de similaridade por medidas de distância no espaço multidimensional dos dados. 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. A distância Euclidiana é definida pela equação 1 (RUAN et al., 2017).

$$Dist(X_1, X_2) = \left(\sum_{i=1}^{k}(X_{1i} - X_{2i})^p\right)^{1/p} \quad (1)$$

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*. Em relação a medição da distância existe diferentes métodos como Mahalanobis, Manhattan, Minkowski, Chebychev, Euclidiana padrão, entre outros. O método varia de acordo com o parâmetro p, onde p=1 é o cálculo da distância de Manhattan, p=2 é a distância de Minkwski e p= ∞ é o método de Chebychev (VERCELLIS, 2009).

O método K-NN é um modelo que exige um grande esforço computacional, 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 facilidade de implementação e acurácia (RUAN et al., 2017).

3. DESENVOLVIMENTO

Este trabalho propõe a construção do sistema integrado capaz de fornecer o monitoramento e a predição de indicadores de desempenho de um processo industrial. Para isso ele 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.

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 isso foi necessário a construção de toda a arquitetura de um sistema de BI, conforme mostrado na figura 1.

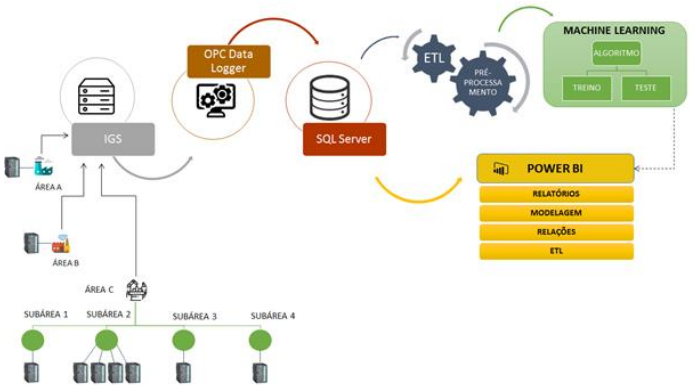

Fig. 1 Esquema geral do Sistema de Predição Industrial

De acordo com a figura 1, o sistema de BI proposto se inicia na coleta de dados do chão de fábrica das diferentes áreas da empresa. Então é realizado o tratamento, filtragem e armazenamento dos dados. Em seguida é realizado as consultas ao banco de dados para confecção das telas e relatórios destinados à gerência da organização. Por fim é feita a avaliação e proposta de melhoria do projeto. Cada uma destas etapas será detalhada nas próximas seções.

Coleta de Dados

A primeira 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 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, onde nenhum sensor foi adicionado, de modo que os dados a serem coletados derivam de variáveis de processo já consolidados.

O controle do processo é realizado pelos Controladores Lógico Programáveis (CLPs) modelos S7-300 e S7-400 com o ambiente Step 7, do fabricante Siemens. Para tornar o processo de coleta mais eficiente, tornou-se necessário a preparação de dados da seguinte forma: (1) 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; (2) 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.

Sendo assim, três grandes áreas produtivas foram envolvidas, sendo a área C da Figura 1, a de maior importância para este trabalho. Este projeto foi composto de nove CLPs e uma de quantidade máxima de 450 variáveis coletadas de cada CLP.

Para a coleta de dados do processo para o sistema SCADA, utilizava-se da tecnologia OPC (OLE for Process Control). O OPC é um conjunto de padrões de comunicação de dados para a indústria para sistemas Windows, capaz de conectar objetos de dados em diferentes protocolos. Neste caso, foi escolhido o IGS (Industrial Gateway Server) da General Eletric. O

software IGS 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. Para comunicação com os diferentes CLPs, o IGS utiliza drivers padrão Profinet, organizando os dados coletados em grupos, o que foi fundamental, uma vez que as variáveis escolhidas derivam de várias partes diferentes do processo produtivo.

Para se conectar ao OPC Server foi utilizado o OPC Client Advanced OPC Data Logger da empresa AGG Software. O software foi projetado para coletar dados do campo e 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 (AGG Software, 2019).

Armazenamento de Dados

No contexto das ferramentas de BI, os dados operacionais devem ser armazenados para posteriormente serem tratados e carregados nas análises de usuário.

Uma configuração importante no *Data Logger* é a taxa de coleta dos dados. 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. No caso, foi utilizado compactação com um dead band de 0,1%.

Para o banco de dados (DB), optou-se pelo SQL Server da Microsoft (2019), uma vez que a organização já possuía licença do banco de dados disponível para ser utilizada. Trata-se de um sistema gerenciador de banco de dados relacional, cuja linguagem de consulta primária é o Transact SQL. Para se conectar ao DB utilizou-se a fonte de dados ODBC. Este recurso é acessado no Administrador de Fonte de Dados ODBC do Windows. Então, é possível configurar o OPC Data Logger para enviar os dados para serem armazenados no DB.

Depois que as configurações de ODBC foram estabelecidas, é possível acessar a instância padrão de banco de dados do SQL ou criar uma nova. Essa conexão pode ser local ou remota, utilizando credenciais próprias do DB, recurso bastante útil para acessar o DB em qualquer servidor de uma rede industrial. Após conectado à uma instância do SQL, foi configurado todos os dados dentro do mesmo DB, para facilitar os backups a serem realizados. No entanto, pode-se criar bancos separados por área, ou assunto caso seja necessário.

Para criação das tabelas contendo os dados operacionais deste trabalho, foi necessário realizar um levantamento de todos 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 2.

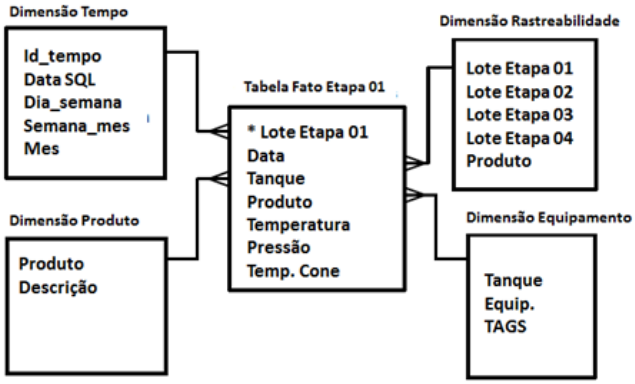

Fig. 2 Exemplo de modelagem utilizada no DB.

Monitoramento e Análise

Para o Monitoramento e análise dos dados foi escolhido o software Power BI da empresa Microsoft. O Power BI é líder de mercado nas soluções em BI e foi escolhido por esta razão para o desenvolvimento deste trabalho. O Power BI permite conectar-se com fontes de dados operacionais diversas, como: Excel, SQL Server, XML, entre outros. Neste projeto, foram utilizados o SQL Server e planilhas Excel como fonte de dados (MICROSOFT, 2019).

A conexão do Power BI pode ser com o *dataset* interno, ou através de *Direct Query*, o qual os dados ficam no DB de origem, e conforme as manipulações ocorrem, as consultas vão sendo executadas. 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, através de consultas.

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. Vale ressaltar a importância da correta configuração das tabelas e

relacionamento entre elas, para garantir a rastreabilidade dos lotes do produto em cada etapa do processo.

A partir do tratamento dos dados, é possível criar telas de visualização como gráficos, indicadores, visualização de KPIs, cartões com exibição de média, desvio padrão, entre outros. As figuras 4 e 5 mostram exemplo de tela de relatórios criados 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.

Predições de Desvios

Após a consolidação das informações do processo através do armazenamento em banco de dados, iniciou-se a etapa de estudos de modelos de predições que agregassem valor para a tomada de decisão da empresa. Para tal, foram utilizados os conceitos de modelos preditivos de ML. 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 para a empresa. Desta maneira, as soluções apresentadas nesse trabalho compõem-se de um protótipo inicial para implementação efetiva do ML nos processos da indústria.

Um algoritmo de ML 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. 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.

Desta forma, para este trabalho foi escolhido o método K-NN de ML pela sua facilidade de implementação e por já possuir bibliotecas padrões em python, que foram os principais critérios para os experimentos iniciais dentro da empresa.

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“ (DUA&GRAFF, 2017).

Para o desenvolvimento dos algoritmos apresentados nesse trabalho foi utilizada a linguagem Python com o pacote Pandas, pela sua facilidade e as bibliotecas de ML existentes, que oferece todo suporte para análises e manipulação de dados. Ainda dentro da linguagem em Python, optou-se também pela utilização da biblioteca Scikit-Learn, com base fundamental para realização dos algoritmos de aprendizado. De maneira simplificada, a biblioteca Pandas permite a manipulação de *Data Frames*, 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, tratamento de vetores, aplicação de filtros, cálculos de média, desvio padrão, aplicação de índice automático, entre outros. Ela também 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, o que foi mais um motivo para adoção desta plataforma neste projeto (GÉRON, 2017).

Conforme apresentado, os algoritmos de classificação fazem parte dos modelos preditivos, e são caracterizados por conterem entradas com rótulos atribuídos a 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 3.

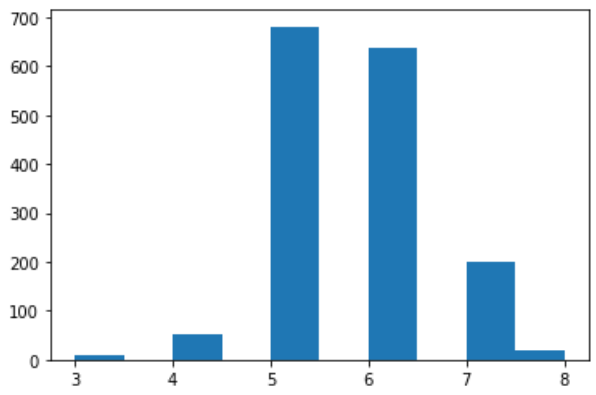

Fig. 3 Histograma dos dados de qualidade.

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 primeira etapa é constituída de realizar o pré-processamento dos dados da seguinte forma: Agrupamento dos dados por lote utilizando o percentual dentro de faixa (valores padrão); Concatenação de tabelas contendo informações de diferentes etapas produtivas; Atribuição de classes como presença e ausência do desvio no produto acabado (0 ou 1); Normalização e padronização de dados.

Após realizado o pré-processamento, foi aplicado o modelo de classificação K-NN, 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.

4. RESULTADOS

Nesta seção serão apresentados os resultados obtidos por este projeto dentro da empresa em estudo. Este trabalho propõe a construção do sistema integrado de BI através da ferramenta de monitoramento e análise e também de uma ferramenta de ML de predição.

Ferramenta de Monitoramento e Análise

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.

A estrutura para o fluxo de dados, desde sua coleta nos CLPs, passando pelo servidor e cliente OPC e sendo armazenado no DB SQL Server mostrou-se efetiva e sem nenhuma falha capaz 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.

Para mostrar os resultados obtidos, aqui será mostrado um dos casos estudados que é 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.

Após a escolha das variáveis, foi identificado na automação 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 diretamente das receitas. Então, foi criado uma lógica no CLP 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 processo.

No sistema SCADA, foi realizada a coleta de dados via OPC com tempo de coleta de 500ms, enquanto que no OPC Data Logger foi utilizado um tempo de 1s, visando garantir uma boa sincronia de envio e recebimento de dados. Então no DB 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. Para o tratamento inicial foram excluídos os outliers, zeros e nulos do SQL pois eles podem indicar um dado incorreto, feito através de uma falha de supervisão, por exemplo e eles não teriam significado estático e poderia trazer alterações nos resultados. Porém, existia a preocupação de medir a porcentagem de dados descartados que existia em um determinado conjunto de amostras, pois isso poderia significar problema de medição.

Finalmente, com a conexão configurada com o DB, 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 a esses parâmetros. A figura 4 mostra o relatório criado.

Fig. 4 Histograma dos dados de qualidade.

De acordo com a figura 4, a cor preta representa o valor do *setpoint* da variável, enquanto que a vermelha apresenta os valores máximo e mínimos 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 10ºC para 15ºC 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 5 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.

Fig. 5 Histograma dos dados de qualidade.

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.

O sistema descrito criado permitiu maior velocidade e facilidade na identificação de problemas em área, como por exemplo a indicação pela ferramenta de variáveis de processo fora dos valores padrão, que ao buscar a resposta para tais divergências, descobriu-se ser falhas causadas em dispositivos por desgaste mecânico.

Uma conclusão que pode ser obtida neste estudo é que o sistema desenvolvido combinada com as metodologias dos 5 Porquês e PDCA é uma importante ferramenta para identificação de problemas e melhoria da qualidade.

Ferramenta de Predição

Em relação a ferramenta de Predição, foram utilizados dois conjuntos de dados: didático e real. Nos testes didático foi utilizado o conjunto de dados disponível em Dua & Graff (2017). Para o teste real foi utilizado dados reais da indústria analisado neste trabalho.

No algoritmo K-NN foi utilizado o método Minkowski para cálculo da distância utilizando p=2 em (1), com número de vizinhos igual a 13. O conjunto de treinamento (k) utilizado foi de 1600 amostras para o DB didático e 123 amostra para os DB real. Para determinar o valor de K foi utilizado o método GridSearchCV da biblioteca sklearn, o qual permite identificar qual o melhor K para um determinado range especificado.

Em relação ao caso real, considerou-se o desvio a presença de alteração sensorial típica no produto final relacionada com a etapa de fermentação do produto, na qual as variáveis de controle são a temperatura, contrapressão e incorporação de oxigênio. Desta forma, as entradas do algoritmo representam o valor total dentro da faixa esperada para um mesmo lote, com tolerância também determinada pela receita da marca.

Estudos preliminares mostraram 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 1 mostra o comparativo entre os resultados obtidos entre os dois conjuntos de dados.

Tabela 1. 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

A Matriz de confusão é uma tabela que categoriza as frequências de classificação do modelo em Verdadeiro Positivo (quadrante superior esquerdo), Falso Positivo (quadrante superior direito), Falso Negativo (quadrante inferior esquerdo) e Verdadeiro Negativo (quadrante inferior direito). A figura 6 mostra a matriz de confusão dos dados obtidos da predição para um estudo didático e também aplicado no caso real.

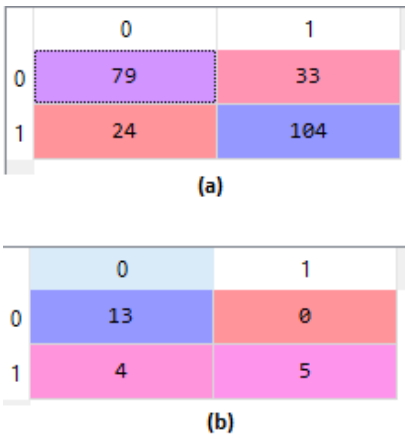

Fig. 6 Matriz de Confusão dos Dados**: (a)** Qualidade (didático)**, (b)** Processo Real

De acordo com a figura 6, observando os resultados do caso real, nota-se uma excelente precisão, de modo que todos os casos classificados positivos para a presença de desvios no

produto acabado realmente apresentam tal desvio. Por outro lado, nos casos classificados como ausência de desvio no produto acabado a taxa de acerto foi reduzida, apresentando uma baixa sensibilidade do modelo, sendo então uma desvantagem estratégica, umas vezes que a previsão tem perdas de previsibilidade de defeitos realmente existentes.

De acordo com a análise das tabelas 1 e da matriz de confusão, 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. CONCLUSÕES

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 supervisório. 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. A partir da conclusão de um modelo mais assertivo, a predição será incorporada ao dashboard, como forma de alertar a possibilidade de ocorrência de algum desvio no produto final, de acordo com o andamento do processo e comportamento das variáveis de controle.

AGRADECIMENTOS

Os autores agradecem o apoio acadêmico e à Estrutura de Pesquisa da Faculdade de Engenharia Elétrica da UFU (Universidade Federal de Uberlândia).

REFERÊNCIAS

AGG SOFTWARE (2019) - The OPC UA Server plugin PRINTED MANUAL, Disponível em: //www.aggsoft.com/opc-data-logger, Acesso em: 06 de maio de 2020.

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

BOSE, I., & MAHAPATRA, R. K. - Business data mining – a machine learning perspective, Information & Management, vol 39, number 3, pages 211 – 225, 2001.

CHEN, C., YAN, X., ZHU, F., HAN, J. YU, P. S. - Graph OLAP: Towards Online Analytical Processing on Graphs∗, 2008 Eighth IEEE International Conference on Data Mining.

DUA, D. and GRAFF, C - UCI Machine Learning Repository, University of California, Irvine, School of Information and Computer Sciences, Disponível em: <https://archive.ics.uci.edu/ml >, Acesso em: Dez. 2019

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

MICROSOFT (2019) - Power BI, Disponível em: https://docs.microsoft.com/pt-br/power-bi/, Acesso em : 06 de maio de 2020

RUAN, Y., XUE, X., LIU, H. TAN, J. LI, Xi - Quantum Algorithm for K-Nearest Neighbors Classification Based on the Metric of Hamming Distance, International Journal of Theoretical Physics, 2017

SAHAY B.S., and RANJAN, J. - Real time business intelligence in supply chain analytics, Information Management & Computer Security, Vol. 16 No. 1, 2008 pp. 28-48.

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

SZEWRAŃSKI S., KAZAK J., SYLLA M., ŚWIĄDER M. (2017) Spatial Data Analysis with the Use of ArcGIS and Tableau Systems. In: Ivan I., Singleton A., Horák J., Inspektor T. (eds) The Rise of Big Spatial Data. Lecture Notes in Geoinformation and Cartography. Springer, Cham

TAN, P.-N., STEINBACH, M., KARPATNE, A. & KUMAR, V. - 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.