MÉTODOS DE MACHINE LEARNING PARA CLASSIFICAÇÃO DA TEMPERATURA NO PROCESSO DE MISTURA EM UMA PLANTA INDUSTRIAL

Glaucia Maria Bressan (UTFPR – Cornélio Procópio) E-mail: glauciabressan@utfpr.edu.br
Guilherme da Cunha (UTFPR – Cornélio Procópio) E-mail: guilherme_cunha20@hotmail.com
Wagner Endo (UTFPR – Cornélio Procópio) E-mail: wendo@utfpr.edu.br

Resumo: O objetivo deste trabalho é a aplicação e análise de métodos de *Machine Learning* para a classificação da temperatura resultante do processo de mistura de líquidos de uma planta industrial didática localizada na UTFPR do câmpus Cornélio Procópio. Os métodos aplicáveis para esta classificação são k-nearest neighbors (KNN), Árvores de Decisão, Florestas Randômicas e *Naive- Bayes*. Para esta tarefa, as variáveis de entrada consideradas são a porcentagem de abertura da válvula, a vazão e o tempo da abertura; e a variável de saída é a temperatura, discretizada em 5 classes. O desempenho dos algoritmos é analisado considerando-se a acurácia e medidas estatísticas relevantes e as implementações dos métodos são feitas utilizando o *software* R. Os resultados apresentam um bom desempenho dos algoritmos na tarefa de classificação da temperatura do processo de mistura de líquidos, com acurácias acima de 90%.

Palavras-chave: algoritmos de *machine learning*, aquecimento, temperatura, processo de mistura.

MACHINE LEARNING METHODS FOR CLASSIFICATION OF TEMPERATURE IN THE MIXING PROCESS IN AN INDUSTRIAL PLANT

Abstract: The goal of this paper is the application and analysis of Machine Learning methods to classify the temperature resulting from the liquids mixing process in a didactic industrial plant located in the UTFPR of the Cornélio Procópio city. The applicable methods for this classification are k-nearest neighbors (KNN), Decision Trees, Random Forest and Naive- Bayes. For this task, the input variables considered are the percentage of valve opening, the flow and the opening time; and the output variable is temperature, discretized into 5 classes. The performance of the algorithms is analyzed considering the accuracy and relevant statistical measures and the implementations of the methods are made using the R software. The results show a good performance of the algorithms in the task of classifying the temperature of the liquid mixing process, with accuracy above 90%.

Keywords: machine learning process, heating, temperature, mixing process

1. Introdução

Os métodos de *Machine Learning*, ou "aprendizado de máquina", são métodos de análise de dados provenientes do estudo da inteligência artificial e têm como principal objetivo automatizar a construção de modelos analíticos. A ideia é que o sistema possa analisar os dados conhecidos, identificando padrões, e com isso, possa tomar decisões sem a intervenção humana (AGGARWAL, 2015).

Atualmente, os métodos de *Machine Learning* são utilizados em diversas aplicações que envolvem classificação ou regressão, a partir do treinamento de um conjunto de dados para reconhecimento de padrões (BISHOP, 2011; AGGARWAL, 2015). Neste trabalho, são aplicados diferentes métodos de *Machine Learning* para análise desempenho dos algoritmos na tarefa de classificar a temperatura no processo de mistura de líquidos de uma planta industrial didática localizada na Universidade Tecnológica Federal do Paraná, campus Cornélio Procópio.

O processo em estudo consiste na mistura de líquidos, que ocorre a partir do líquido proveniente do tanque de aquecimento e do líquido proveniente do reservatório de água fria. O último responde às variações no tanque de aquecimento, a partir de seus controladores, mantendo a temperatura no valor desejado. A ação do controlador se dá por uma relação entre a demanda de entrada da água fria e a temperatura que é medida no tanque de aquecimento (SILVA, ENDO e LISBOA, 2011). É importante ressaltar que, conforme se aumenta a vazão, a temperatura da mistura também aumenta. Além disso, quando a válvula é acionada diversas vezes em sequência, nota-se resquícios desse aumento de temperatura na mistura. Tal justifica a consideração do tempo no processo de mistura e da implementação dos métodos de aprendizado de máquina para a classificação da temperatura.

O trabalho de Bressan, Mosaner e Endo (2020) propõe o estudo e aplicação de uma estratégia unificada de classificação automática em uma malha de controle de vazão, com uma válvula pneumática industrial como elemento final de controle, utilizando a mesma planta industrial em estudo. Os autores elaboraram um sistema de classificação fuzzy e a saída do sistema desenvolvido é compensado diretamente na ação de controle na temperatura do líquido de um sistema de aquecimento.

Neste contexto, o objetivo deste trabalho é a aplicação e a análise de métodos de *Machine Learning* para a classificação da temperatura resultante do processo de mistura de líquidos de uma planta industrial didática localizada na UTFPR do câmpus Cornélio Procópio. Os métodos considerados para esta classificação são KNN, Árvores de Decisão, Florestas Randômicas e *Naive- Bayes*, descritos detalhadamente na Seção 3. Para isso, as variáveis de entrada consideradas são a porcentagem de abertura da válvula, a vazão e o tempo da abertura; e a variável de saída é a temperatura, a qual é discretizada em 5 classes (muito baixa, baixa, média, alta e muito alta). O desempenho dos algoritmos é analisado considerando-se a acurácia e medidas estatísticas relevantes, e as implementações dos algoritmos são feitas utilizando o *software* R (https://cran.r-project.org/bin/windows/base/).

Este trabalho está organizado da seguinte forma: na Seção 2 é apresentada a descrição da planta industrial didática e a instrumentação física dos processos industriais. Na Seção 3 são apresentados os algoritmos de *Machine Learning* utilizados para a classificação da temperatura proveniente do processo de mistura de líquidos. A Seção 4 apresenta os resultados dos classificadores e suas respectivas análises estatísticas. Na Seção 4, as considerações e sugestões para futuros trabalhos são descritas.

2. Descrição da Planta Industrial Didática

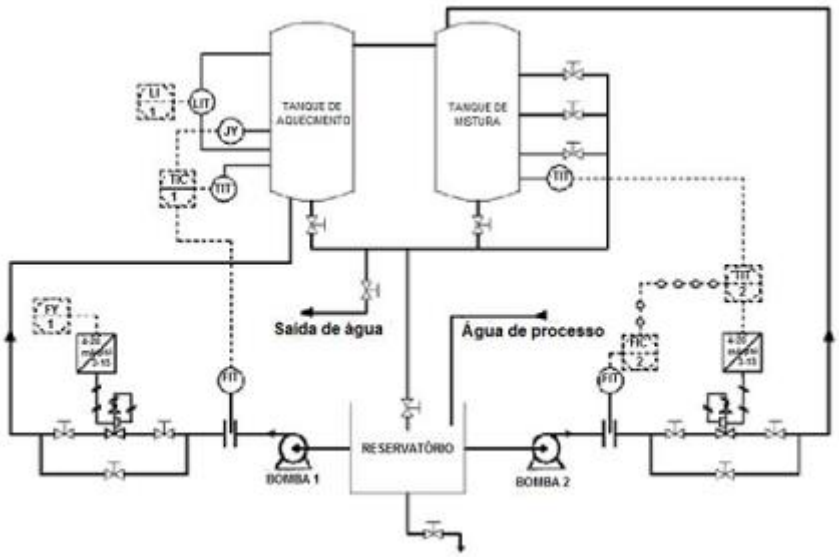
Este trabalho foi desenvolvido utilizando a planta didática SMAR, ilustrada na Figura 1, que reproduz processos industriais para fins didáticos, podendo simular experimentos muito próximo de situações reais.

Figura 1 – Planta industrial didática SMAR

A planta didática simula dois processos, sendo eles, aquecimento e mistura. O aquecimento ocorre através de uma resistência de submersão, instalada no tanque de aquecimento. Já a mistura é feita a partir do líquido aquecido e do frio, proveniente do outro reservatório. Este é acionado a partir de um controlador, e responde as variações de temperatura do tanque quente, mantendo a mistura na temperatura desejada (*setada* previamente).

O sistema identificará a demanda de água fria, e a utilizará para manter a temperatura desejada, a misturando com a água quente. O processo procura o equilíbrio térmico desejado de forma empírica. A Figura 2 ilustra o diagrama de tubulação e instrumentação da planta estudada.

Figura 2 – Diagrama de tubulação e instrumentação



As siglas de instrumentação presentes no diagrama da planta são descritas como segue.

LI: Indicador de nível
LIT: Transmissor indicador de nível
JY: Conversor de potencia
TIT: Transmissor indicador de temperatura
TIC: Controlador indicador de temperatura

FY: Conversor de vazão
FIT: Transmissor indicador de vazão

A aquisição de dados dos equipamentos é feita por meio do sistema supervisório ProcessView integrado ao SYSTEM302, além disso, este possui uma interface gráfica que mostra as informações presentes na malha de controle. A utilização do software Matlab como cliente OPC permite a coleta de dados do processo através de uma interface de aquisição de dados.

3. Metodologia

Nesta seção, são apresentados os métodos de *Machine Learning* utilizados neste trabalho para a tarefa de classificação da temperatura, bem como os métodos de validação dos resultados.

Optou-se por utilizar o *software* R para implementação dos classificadores, visto que este já possui os pacotes necessários. Primeiramente, é necessário realizar o pré-processamento dos dados, por meio do tratamento de dados faltantes e *outliers*, e selecionando as variáveis de entrada do problema. É importante notar que o pré-processamento é sempre o primeiro passo a ser realizado, independente do modelo escolhido para realizar a tarefa de classificação.

Após realizado o pré-processamento dos dados, pode-se iniciar a tarefa de classificação por meio dos algoritmos de *Machine Learning*. Neste trabalho, os modelos aplicáveis para a classificação da temperatura no processo de mistura na planta industrial são: KNN, Árvores de Decisão, Florestas Randômicas e *Naive- Bayes*, descritos a seguir.

3.1. O Método KNN

O método k-*Nearest Neighbors* (KNN), ou "K vizinhos mais próximos", é um dos classificadores mais conhecidos, e costuma ser muito utilizado por conta da facilidade de compreensão e implementação. O objetivo é determinar o rótulo da classificação de uma amostra com base nas amostras vizinhas advindas dos dados de treinamento.

Para o bom funcionamento do algoritmo, existem dois pontos principais que devem ser analisados: O valor de *k* e a medida de similaridade, ou seja, a métrica de distância para encontrar os vizinhos mais próximos. Para a medida de similaridade, a distância Euclidiana é sem dúvidas a medida mais utilizada (AGGARWAL, 2015).

Em relação ao parâmetro *k*, existem diversas alternativas para determina-lo É possível utilizar um algoritmo de otimização para encontrar o melhor valor para o conjunto de dados, entretanto, o desempenho geral do modelo será lento na etapa de seleção do *k*. Outra forma comumente utilizada é avaliar o algoritmo no conjunto de validação, testando diferentes valores d e*k*, ou seja, encontrar o melhor valor de *k* empiricamente.

3.2. O Método das Árvores de Decisão

As árvores de decisão estão entre os modelos mais utilizados em *Machine Learning*, principalmente em problemas de classificação. Tais algoritmos estão fundamentadas no paradigma *Bottom-up*, ou seja, a obtenção do modelo de classificação é feita pela identificação de relacionamentos entre variáveis dependentes e independentes (QUINLAN, 2014). Diferentemente do algoritmo KNN, nas árvores de decisão o conhecimento adquirido pode ser representado por meio de regras, as quais podem ser expressas em linguagem neural, facilitando sua compreensão.

O chamado *Top-Down induction of Decision Tree* (TDIDT) é um algoritmo conhecido, utilizado como base para diversos algoritmos de árvores de decisão, como ID3

(QUINLAN, 1986; QUINLAN, 1988), C4.5 (QUINLAN, 2014) e CART (BREIMAN et al., 1984). Tal algoritmo produz regras de decisão de forma implícita em uma árvore de decisão. Segundo BRAMER (2007), o processo é conhecido como *particionamento recursivo*, onde a árvore é construída por sucessivas divisões dos exemplos de acordo com os valores de seus atributos preditivos.

Para definir qual atributo preditivo é utilizado em cada nó da árvore, é utilizado um critério de seleção. Tais critérios são definidos em termos da distribuição de classe dos exemplos antes e depois da divisão (TAN et al., 2005).

Normalmente, nos algoritmos de indução de árvores de decisão, cada nó é dividido de acordo com um único atributo. Os critérios de seleção para a melhor divisão são baseados em diferentes medidas, como impureza, distância e dependência. A maior parte dos algoritmos desse tipo busca dividir os dados de um nó-pai de forma a minimizar o grau de impureza dos nós-filhos.

Uma das medidas mais utilizadas é o *ganho de informação*, o qual utiliza *entropia* como medida de impureza. O algoritmo ID3 (QUINLAN, 1986), pioneiro em indução de árvores de decisão utiliza essa medida. É necessário calcular o ganho de informação, comparando o grau de entropia do nó-pai com o grau de entropia dos nós-filhos. O atributo que gera uma maior diferença (um maior ganho de informação) é escolhido como condição de teste. O ganho de informação é dado pela equação (1).

$$G = Entropia\ (pai) - \sum_{j=1}^{n}[\frac{N(v_j)}{N} Entropia(v_j)]\ , \qquad (1)$$

em que *n* é o número de valores do atributo, *N* é o número total de objetos do nó-pai e $N(v_j)$ é o número de exemplos associados ao nó-filho v_j. Já o grau de entropia, é definido pela Equação (2).

$$Entropia(nó) = -\sum_{i=1}^{c} p\left(\frac{i}{nó}\right) \log_2[p(i/nó)] \qquad (2)$$

O termo $p(\frac{i}{nó})$ é a fração dos registros pertencentes à classe *i* no nó, e *c* é o número de classes.

O critério de ganho seleciona como atributo-teste aquele que maximiza o ganho de informação. O grande problema ao se utilizar essa medida é que ela dá preferência a atributos com muitos valores possíveis. Para solucionar tal problema, Quinlan (1993) propôs a razão de ganho como critérios de avaliação. A razão de ganho é dada pela equação (3).

$$razão\ de\ ganho(nó) = \frac{ganho}{entropia(nó)} \qquad (3)$$

Por fim, em Quinlan (1988), é sugerido que a razão de ganho seja calculada em duas etapas. Na primeira, o ganho de informação é calculado para todos os atributos, após isso, considera-se apenas aqueles que obtiveram um ganho acima da média e a partir disso, escolhe-se aquele que possuir a melhor razão de ganho.

Após as árvores de decisão serem construídas, muitas delas apresentam o problema conhecido como sobreajuste (QUINLAN, 1988), em que o aprendizado é muito específico para o conjunto de treinamento, não permitindo ao modelo se generalizar e ter uma boa eficiência nos dados de teste. Para contornar esse problema, são utilizados métodos de poda (*pruning*) da árvore, cujo objetivo é melhorar a taxa de acerto do modelo, removendo-se ramos completos (HAN, 2001). Para decidir se a árvore é

podada, é calculada uma taxa de erro para caso a árvore seja podada, após isso, é calculada a mesma taxa para caso não seja podada. Se a diferença for menor que o valor pré-estabelecido, a árvore é podada.

3.3. O método Florestas randômicas

O método *Randon Forests*, ou Florestas Randômicas, funciona como uma melhoria nos resultados de uma árvore de decisão pois, como o nome sugere, cria diversas alvores de decisão de maneira aleatória para um mesmo problema, onde cada árvore influenciará no resultado final (BREIMAN, 2001).

Tal algoritmo faz parte do conjunto de métodos *ensemble*, tais quais combinam diferentes modelos para se obter um único resultado, tornando os algoritmos mais robustos e complexos, elevando o custo computacional, porém trazendo melhores resultados. (BREIMAN, 2001).

A principal vantagem na utilização desses algoritmos se dá pela combinação dos resultados, visto que normalmente, em um método comum, é escolhido o modelo que apresenta melhores resultados para o conjunto de dados, entretanto, mesmo sendo possível testar diferentes configurações para o modelo escolhido, ao final, é escolhida apenas uma configuração. Com a utilização de um método *ensemble*, todos os resultados referentes a todas as configurações testadas são utilizados, eliminando assim problemas de instabilidade.

3.4. O Método Naive Bayes

O modelo de inferência Bayesiana consiste em analisar variáveis condicionadas de maneira probabilística. Isso pode ser feito por meio do teorema de Bayes: Suponha que y' = $(y_1, \cdots, y_n)$ seja um vetor de *n* observações no qual a distribuição de probabilidade $p(y|\theta)$ dependa do valor de k parâmetros θ' = $(\theta_1, \ldots, \theta_k)$. Considere que θ tenha distribuição de probabilidade p(θ). Assim, dada a observação y, a probabilidade condicional de θ é representada pelo teorema de Bayes (BOX; TIAO, 1992), o qual é mostrado na equação (4).

$$p(\theta|y) = \frac{p(y|\theta).p(\theta)}{p(y)} \quad (4)$$

O termo *p*(θ) é chamado de destruição a priori de θ, $p(\theta|y)$ de distribuição a posteriori de θ dado *y* e $p(y|\theta)$ de v*erossimilhança* de θ.

O maior problema prático é que normalmente em casos reais a variável de interesse depende de outras. Nota-se que é possível expandir o teorema para diversas variáveis, porém isso não é recomendado devido à complexidade dos cálculos (NEAPOLITAN, 2003). Tendo isso em vista, Pearl (1988) desenvolveu as conhecidas redes bayesianas, que avaliam as interligações das variáveis por meio de suas estruturas.

As redes bayesianas são modelos gráficos probabilísticos que representam o conhecimento sobre o domínio dos dados. O conhecimento *a priori* pode ser combinado com padrões aprendidos a partir dos dados, além disso, o usuário pode inserir conhecimento em algum nó da rede, fazendo com que isso se propague aos outros nós. Os modelos são compostos por um grafo acíclico direcionado e um conjunto de tabelas de probabilidade, onde os nós da rede representam as variáveis e os arcos representam relações de dependência entre as variáveis.

As Redes Bayesianas podem ser utilizadas como classificadores, calculando a probabilidade condicional de um nó, chamado nó classe, dados os valores das

probabilidades dos outros nós, ou seja, $P(C|V)$, onde *C* representa a classe analisada e *V* o conjunto de variáveis que descreve os padrões. O classificador bayesiano mais importante é o *Naive Bayes*, que se destaca pelos sucessos obtidos na aplicação de diversos problemas apesar de sua simplicidade, mesmo quando comparado a classificadores mais complexos (FRIEDMAN; GEIGER; GOLDSZMIDT, 1997).

O modelo *Naive Bayes* descreve um caso particular de uma Rede Bayesiana, onde é considerado que as variáveis do domínio são condicionalmente independentes. A classificação é feita aplicando o teorema de Bayes para calcular a probabilidade de *C*, dado uma instância particular de $A_1, A_2, \ldots A_n$ e então predizendo a classe com maior probabilidade *a posteriori* (FRIEDMAN; GEIGER; GOLDSZMIDT, 1997). Então, o processo de aprendizagem é feito de maneira indutiva, apresentando um conjunto de dados de treinamento e calculando a probabilidade condicional de cada atributo A_i, dado a classe *C* (FRIEDMAN; GEIGER; GOLDSZMIDT, 1997).

3.5. Validação dos Resultados

A *validação cruzada* é uma técnica que será empregada ao fim de todos os algoritmos, a fim de aumentar a credibilidade dos métodos. A técnica consiste em dividir os dados de treinamento em *n* subconjuntos distintos. Um dos subconjuntos será usado para teste, enquanto os outros serão usados para treinamento. O processo é repetido usando-se cada um dos *n* subconjuntos como dados de teste, e ao final, o erro será medido sobre todas as possibilidades. Será adotado $n = 10$ por ser o valor mais comumente encontrado na literatura (AGGARWAL,2015).

A *matriz de confusão* é utilizada para uma verificação primária da aplicação de um classificador. Em suma, a matriz contém o número de elementos que foram classificados corretamente e incorretamente. Tais valores são representados em uma matriz, onde a diagonal principal representa os elementos corretamente classificados, enquanto os valores que estão fora dela representam o número de elementos classificados de forma incorreta. A matriz de confusão será plotada para todos os classificadores com intuito de comparação.

4. Resultados e Discussão

Nesta seção, são apresentados os resultados numéricos provenientes dos algoritmos executados, bem como a comparação e a discussão de suas performances na tarefa de classificação de falhas da planta industrial didática em estudo.

4.1. Pré-processamento dos dados

O conjunto de dados contém 1202 linhas, que consistem nos valores numéricos das 3 variáveis de entrada (percentual de abertura da válvula, tempo e vazão) e também o valor numérico de saída (Temperatura). Todos os dados estão normalizados, ou seja, variam de 0 a 1.

Para realizar a classificação, primeiramente, é necessário discretizar a variável de saída, Temperatura, de modo hierárquico. De acordo com Bressan; Mosaner; Endo, (2020), foram escolhidas cinco classes de temperatura: 1- muito baixa, 2- baixa, 3- média, 4- alta e 5- muito alta.

O conjunto de dados é dividido nos subconjuntos de treinamento (treinamento do modelo, que consiste em "aprender" a classificar corretamente a variável de saída) e de teste (utilizado para verificação de sua acurácia, ou índice de acerto da classificação), de acordo com o método de validação cruzada com $n = 10$ (AGGARWAL,2015).

4.2. Resultados do método KNN

O método KNN é implementado no *software* R com o auxílio das bibliotecas *"caTools", "class"* e "*caret*", de forma que seja possível realizar a classificação e obter a matriz de confusão, a fim de verificar o desempenho da classificação. Além disso, os resultados estatísticos auxiliam para a análise do desempenho do método.

A matriz de confusão para o método KNN utilizando um número de vizinhos $k = 9$ é dada a seguir.

classes	**1**	**2**	**3**	**4**	**5**
1	26	0	0	0	0
2	0	20	0	0	0
3	0	0	44	2	0
4	0	0	0	82	2
5	0	0	0	1	124

Vale ressaltar que o número de vizinhos a ser utilizado foi obtido de forma iterativa, simplesmente testando valores comuns de utilização, a começar por *k=5*. O método se mostrou extremamente eficaz para o problema analisado, com apenas 5 classificações erradas e nenhuma classificação discrepante. Como é possível notar resquícios do aumento de temperatura quando o tempo de vazão aumenta, era esperado obter mais erros em classes de temperatura mais alta, mesmo assim, a maioria dos valores encontra-se na diagonal principal da matriz, a qual representa os acertos do método.

A Tabela 1 mostra importantes medidas estatísticas do método, necessárias para verificação da eficácia do KNN como classificador para o problema apresentado.

Tabela 1: Medidas Estatísticas do modelo KNN

Medidas Estatísticas	**Valor**
Acurácia	0,9834
Intervalo de confiança	0,9617; 0,9946
Kappa	0,9767
Sensitividade	1 ,1 ,1 , 0,9647, 0,9841 (5 classes em ordem)
Especificidade	1, 1, 0,9922, 0,9907, 0,9943 (5 classes em ordem)

A principal estatística a ser observada é a acurácia, sendo 0,9834 um valor extremamente alto, representando um índice de acerto da classificação de 98,34%. O alto valor de Kappa mostra uma concordância entre os dados, e os valores próximos a 1 para sensitividade e especificidade provam a capacidade de correta classificação do método. As definições para estas estatísticas apresentadas podem ser encontradas em Aggarwal (2015). A Tabela 2 mostra as principais estatísticas para o método KNN utilizando duas rodadas de validação cruzada.

Tabela 2: Validação cruzada para o KNN

k	**Acurácia**	**Kappa**
5	0.9833605	0.9766693
7	0.9825409	0.9754699
9	0.9858537	0.9801361

Observa-se que mesmo com outros valores de *k*, os valores continuam sendo positivos, provando assim, que o método obteve sucesso e selecionando o valor de $k = 9$ como sendo o melhor a ser utilizado, por apresentar maior acurácia.

4.3. Resultados da árvore de decisão

Para a implementação das Árvores de Decisão, são utilizadas as seguintes bibliotecas: *"caTools", "rpart"*, *"rpart.plot"* e *"caret"*. A matriz de confusão é apresentada a seguir.

classes	**1**	**2**	**3**	**4**	**5**
1	26	0	0	0	0
2	0	20	0	0	0
3	0	0	46	2	0
4	0	0	1	82	1
5	0	0	0	1	124

Ao observar a matriz, nota-se o bom desempenho do classificador, visto que quase todos os valores se encontram na diagonal principal, a qual representa os acertos. Em uma primeira análise, o algoritmo se mostra eficiente mesmo com a dificuldade de se classificar as altas temperaturas provenientes de aquecimentos anteriores. A Tabela 3 mostra as medidas estatísticas, a fim de quantificar o desempenho do modelo.

Tabela 3: Estatísticas do modelo de árvore de decisão

Medidas Estatísticas	**Valor**
Acurácia	0,99
Intervalo de confiança	0,9712; 0,9979
Kappa	0,9861
Sensitividade	1, 1, 0,9787, 0,9880, 0,9920 (5 classes em ordem)
Especificidade	1, 1, 1, 0,9908, 0,9943 (5 classes em ordem)

A acurácia de 99% reforça o resultado analisado na matriz de confusão, provando que o classificador possui um excelente desempenho para o caso analisado. O valor do índice Kappa mostra uma alta concordância e os valores de especificidade e sensitividade próximos a 1 aumentam a confiabilidade da classificação.

4.4. Resultados das Florestas Randômicas

Para a implementação das Florestas Randômicas, foram utilizadas as seguintes bibliotecas: *"caTools", "randomForest", "caret".* Como citado anteriormente, esse algoritmo permite uma confiabilidade maior do que as árvores de decisão, pois elimina ou ao menos diminui significativamente a chance de as estatísticas serem influenciadas por seleções boas ou ruins de conjunto de treinamento e teste. A matriz de confusão obtida para o método das florestas é mostrada a seguir.

classes	**1**	**2**	**3**	**4**	**5**
1	26	0	0	0	0
2	0	20	0	0	0
3	0	1	45	0	0
4	0	0	1	77	6
5	0	0	0	1	124

O desempenho do classificador pode ser observado na matriz de confusão, visto que a maioria dos valores se encontram na diagonal principal, que representa os acertos da classificação. A Tabela 4 mostra as medidas estatísticas do método.

Tabela 4: Medidas Estatísticas das Florestas Randômicas

Medidas Estatísticas	**Valor**
Acurácia	0,9701
Intervalo de confiança	0,944; 0,986
Kappa	0,9581
Sensitividade	1 , 0,95238 , 0,9783 , 0,9872, 0,9538 (5 classes em ordem)
Especificidade	1, 1, 0,9961, 0,9686, 0,9942 (5 classes em ordem)

O alto desempenho do classificador pode ser notado pelos valores de acurácia e intervalo de confiança. A acurácia é um pouco menor do que o método das Árvores de Decisão, justamente porque o modelo utiliza uma combinação de 100 árvores para obter tais resultados, o que garante maior confiabilidade do método. O valor do índice Kappa mostra uma boa concordância e a sensitividade e especificidade próximos a 1 aumentam a confiabilidade. Por fim, foram realizadas duas rodadas de validação cruzada para o método. O retorno foi de uma acurácia de 0,9929200 e Kappa de 0,9900797, confirmando o bom desempenho do classificador.

4.5. Resultados do método Naive- Bayes

Para a implementação do classificador *Naive Bayes*, são utilizadas as seguintes bibliotecas: *"caTools", "e1071"* e *"caret"*. Dessa forma, verifica-se os resultados da classificação com o auxílio da matriz de confusão e das estatísticas descritivas. A matriz de confusão é apresentada a seguir.

classes	**1**	**2**	**3**	**4**	**5**
1	26	0	0	0	0
2	0	20	0	0	0
3	0	5	39	2	0
4	0	0	0	74	10
5	0	0	0	11	114

A maioria dos valores se encontra na diagonal principal, mostrando um bom desempenho do classificador, além disso, não há classificações muito discrepantes. Novamente, a maior dificuldade se encontra em classificar as temperaturas mais altas, visto que há resquícios dos aquecimentos anteriores, dificultando a caracterização do comportamento. A Tabela 5 mostra as medidas estatísticas do método, necessárias para verificação da eficácia do *Naive Bayes* como classificador para o problema apresentado.

Tabela 5: Medidas Estatísticas do modelo *Naive Bayes*

Estatística	**Valor**
Acurácia	0,907
Intervalo de confiança	0,8684; 0,9373
Kappa	0,87
Sensitividade	1 , 0,8 ,1 , 0,8506, 0,9194 (5 classes em ordem)
Especificidade	1, 1, 0,9733, 0,9533, 0,9379 (5 classes em ordem)

A acurácia de 90,7% mostra o bom desempenho do método *Naive Bayes*, apesar de outros classificadores citados se mostrarem ainda melhores. O índice Kappa de 0,87 mostra uma boa concordância e os valores de sensitividade e especificidade acima de 0,8 mostram que o classificador tem em geral, uma boa confiabilidade.

Para gerar uma maior confiabilidade e evitar medidas ao acaso, utilizou-se duas rodadas de validação cruzada, obtendo-se uma acurácia de 0,9433408 e valor do índice Kappa de 0,9201878. Isso mostra que as condições do primeiro teste (por exemplo a seleção de dados de treinamento) não era tão favorável, e que em geral, o método tem um desempenho melhor do que o mostrado anteriormente.

4.6. Discussão dos Resultados

Comparando os resultados das estatísticas descritivas, juntamente com as matrizes de confusão de cada método e considerando as validações cruzadas, nota-se primeiramente que o classificador *Naive Bayes* apresentou pior desempenho na classificação em comparação aos outros métodos para o caso apresentado. Isso pode ser justificado pela simplicidade do método e baixíssimo custo computacional envolvido no processamento do mesmo. É importante ressaltar que apesar de comparativamente os resultados para o método *Naive Bayes* serem os piores, em geral, a acurácia de aproximadamente 94% é alta, mostrando que a ferramenta possui um bom custo-benefício.

Em relação à acurácia, os maiores valores (99% ou acima) foram obtidos pelo método das Árvores de Decisão e Florestas Randômicas, entretanto, a acurácia de 98,58% obtida pelo KNN também deve ser notada, visto que o custo computacional envolvido é menor. As Florestas Randômicas apresentaram o maior valor de Kappa, mostrando uma melhor concordância entre os dados, porém, ambos os métodos KNN e Árvores de Decisão também apresentaram valores elevados (acima de 98%). Vale lembrar também que as Florestas Randômicas apresentam resultados mais confiáveis, entretanto, o custo computacional envolvido na criação das florestas é extremamente maior.

Além disso, os valores de sensitividade e especificidade mostram que os métodos podem classificar adequadamente a temperatura, entretanto, os melhores valores são apresentados pelos métodos KNN e Árvores de Decisão, especialmente quando observados os valores das classes mais altas de temperatura, onde a dificuldade de classificar é maior.

Em especial, o método KNN se destaca pela sua constância, mantendo valores extremamente elevados de acurácia, intervalo de confiança, Kappa, sensitividade e especificidade à um custo computacional relativamente baixo quando comparado com as Árvores de Decisão, que apresenta os melhores valores brutos das estatísticas descritivas, porém, a um custo computacional maior e menor confiabilidade e robustez, quando comparado às Florestas Randômicas.

É importante ressaltar que as Árvores de Decisão têm a vantagem na questão da interpretação dos resultados, apresentando um resultado visual ao usuário que pode ser visto como decisões lógicas, facilmente verificadas. Vale ressaltar que o alto custo das Florestas Randômicas e Árvores de Decisão está associado ao processo de criação das árvores, porém, uma vez criadas, o custo é baixo para a utilização dos métodos.

5. Conclusão

Este trabalho propõe aplicação e análise de métodos de *Machine Learning* para a classificação da temperatura resultante do processo de mistura de líquidos de uma planta industrial didática, localizada na UTFPR do câmpus Cornélio Procópio. Todos os

métodos considerados apresentaram bons resultados em relação à acurácia e às medidas estatísticas. O método *Naive Bayes* tem bons resultados a um menor custo computacional, enquanto as Árvores de Decisão mostram as melhores estatísticas descritivas e maior quantidade de acertos a um alto custo computacional. Ao mesmo tempo, as Florestas Randômicas apresentam a maior confiabilidade e robustez, porém, é o método com maior custo computacional, em relação aos demais. Portanto, os classificadores apresentados se mostram como sendo ferramentas eficazes para previsão de dados e com ótimo desempenho associado à tarefa de classificação.

Como perspectivas de continuidade do trabalho, propõe-se uma nova coleta de dados na planta industrial didática, para elaboração de métodos de *Machine Learning* para a classificação de falhas dos processos executados pela planta em estudo. Tais falhas podem ser classificadas em determinados grupos e o desempenho dos algoritmos podem ser analisados de forma similar.

Referências

AGGARWAL, C. C. (Ed.). *Data Classification: algorithms and applications* Chapman & Hall/CRC Data Mining and Knowledge Discovery Series Book 35, 2015.

BISHOP, C. M. *Pattern Recognition and Machine Learning***.** Springer, 2011.

BOX, G.E.P., TIAO, G. C. *Bayesian inference in statistical analysis*. John Wiley and Sons, Canada, 1992.

BRAMER, M. *Principles of data mining*. Springer, London, 2007.

BREIMAN, L., FRIEDMAN, J. H., OLSHEN, R. A., STONE, C. J. *Classification and Regression Trees*. Wadsworth, 1984.

BREIMAN, L. *Statistical modeling: the two cultures*. Statistical Science, vol. 16, n. 3, p.199--231, 2001.

BRESSAN, G. M.; SILVA, G. M.; ENDO, W. *Estratégia para Compensação de Erros de Ação de Controle em uma Válvula Industrial Utilizando Inferência Fuzzy*. Revista de Engenharia e Tecnologia. v.12, p.223 - 234, 2020.

FRIEDMAN, N.; GEIGER, D.; GOLDSZMIDT, M. *Bayesian Network Classifiers*, Machine learning, v. 29, n. 2, p.131--163, 1997.

HAN, J. *Data Mining: concepts and techniques*. Morgan Kaufmann Publishers Inc., San Francisco, CA, 2001.

NEAPOLITAN, R. E. *Learning Bayesian Networks*. Prentice Hall, 2003.

PEARL, J. *Probabilistic Reasoning in Intelligent Systems: networks of plausible inference*. Morgan Kaufmann, San Francisco, CA, 1988.

QUINLAN, J. R. *Induction of Decision Trees*. Machine Learning, v. 1, n. 1, p. 81--106, 1986.

QUINLAN, J.R. *Decision trees and multivalued attributes*. Machine Intelligence 11, p. 305—318, 1988.

QUINLAN, J. R. *C4.5: Programs for machine learning*. Morgan Kaufmann. Publishers Inc., San Francisco, CA, 2014.

SILVA, L. R. B.; ENDO, W. & LISBÔA, A. R. B. S. *Expectativas da utilização de uma planta didática industrial como objeto de aprendizagem em um curso de graduação em engenharia*. *In:* CONGRESSO BRASILEIRO DE EDUCAÇÃO EM ENGENHARIA, 39., 2011, Blumenau. Anais... Blumenau. ABENGE, 2011

TAN, P.N., STEINBACH, M., KUMAR, V. *Introduction to Data Mining*, Addison-Wesley Longman Publishing Co., Inc., Boston, MA, 2005.