Uso do CTGAN na Geração de Dados Sintéticos para Planejamento, Controle da Produção e Predição com Machine Learning

Felipe Miguel 1, Ederson Fernandes 2
1. Estudante do curso de Ciência de Dados na UNINTER
2. Professor e pesquisador na UNINTER

Grupo de trabalho: GT 07 - ENGENHARIA, TECNOLOGIA E INOVAÇÃO

RESUMO

A qualidade dos dados é um fator essencial para garantir a eficiência das decisões tomadas em ambientes produtivos. No contexto do Planejamento e Controle da Produção (PCP), a ausência ou limitação de dados digitais pode comprometer a análise de desempenho, dificultar a identificação de gargalos e limitar a aplicação de soluções baseadas em inteligência artificial. Este trabalho tem como objetivo demonstrar a geração de dados sintéticos por meio do modelo CTGAN (*Conditional Tabular Generative Adversarial Network* – Rede Generativa Adversária Condicional para Tabelas) visando suprir a escassez de dados históricos em ambientes industriais e possibilitar o treinamento de modelos preditivos com maior confiabilidade. Utilizou-se como base o conjunto de dados Predicting Manufacturing Defects Dataset, processado com a biblioteca SDV (*Synthetic Data Vault* – Cofre de Dados Sintéticos). O modelo foi treinado com 3000 épocas e utilizado para gerar um novo conjunto contendo 3240 amostras sintéticas. A avaliação foi realizada com ferramentas como o QualityReport e o DiagnosticReport, além de análises visuais de variáveis categóricas e numéricas. Os resultados demonstraram boa fidelidade entre os dados reais e sintéticos, com pontuação global de qualidade de 0.9185, sendo 0.9037 para formas de colunas e 0.9335 para tendências entre pares de colunas. Conclui-se que o uso de dados sintéticos com CTGAN pode ser uma alternativa eficaz para preencher lacunas de dados em ambientes industriais, apoiando a análise de processos e a implementação de soluções baseadas em aprendizado de máquina.

Palavras-chave: Dados sintéticos; Aprendizado de máquina; Produção industrial.

[Digite aqui]

INTRODUÇÃO

A tomada de decisões eficazes em ambientes produtivos depende diretamente da qualidade e disponibilidade dos dados utilizados nas etapas de planejamento, controle e análise de desempenho. No entanto, muitas empresas ainda enfrentam dificuldades devido à ausência de dados digitalizados ou à limitação de registros históricos consistentes, o que compromete a eficiência de processos e impede a aplicação de soluções baseadas em inteligência artificial. A aplicação de inteligência de dados no setor industrial tem impulsionado transformações profundas, com destaque para a automação, otimização de processos e uso estratégico das informações (MIT TECHNOLOGY REVIEW INSIGHTS, 2023).

Nesse cenário, a geração de dados sintéticos surge como uma alternativa promissora, permitindo preencher lacunas e criar conjuntos de dados realistas a partir de padrões aprendidos com dados reais. Entre os modelos utilizados para esse fim, destaca-se o CTGAN, capaz de lidar com variáveis mistas e gerar amostras com distribuições semelhantes às do conjunto original.

Este trabalho tem como objetivo aplicar o CTGAN na geração de dados sintéticos com base em um conjunto real referente à identificação de defeitos em processos de manufatura, demonstrando sua viabilidade como ferramenta para apoiar o planejamento e controle da produção em contextos com escassez de dados.

METODOLOGIA

Neste trabalho, foi utilizado um conjunto de dados da área industrial para avaliar a geração de dados sintéticos por meio de modelos generativos. Inicialmente, foram importadas bibliotecas essenciais para análise, manipulação e visualização dos dados. Após o carregamento e inspeção do conjunto original, os metadados foram detectados automaticamente.

Em seguida, um modelo foi treinado para gerar um novo conjunto de dados sintéticos com a mesma estrutura do original. Foram analisadas as perdas durante o processo de treinamento e gerados relatórios de diagnóstico e qualidade para avaliar a consistência dos dados gerados.

A semelhança entre os dados reais e sintéticos foi validada por meio de comparações gráficas de distribuições, análise de correlação e uma métrica que avalia o quanto os dados sintéticos conseguem se confundir com os reais.

O desenvolvimento deste trabalho encontra-se em andamento, com futuras etapas previstas para aplicação da metodologia em dados reais de produção industrial e o uso de técnicas de machine learning para prever variáveis relevantes no planejamento e controle da produção.

RESULTADOS E DISCUSSÃO

A geração dos dados sintéticos por meio do CTGAN apresentou resultados satisfatórios quanto à fidelidade estatística e à preservação das características do conjunto original. A avaliação realizada pelo QualityReport indicou uma pontuação global de qualidade de aproximadamente 0,91, com destaque para as propriedades de Column Shapes (0,90) e Column Pair Trends (0,93). Isso demonstra que o modelo foi capaz de reproduzir bem tanto as distribuições individuais quanto as relações entre variáveis.

A comparação entre as matrizes de correlação dos dados reais e sintéticos resultou em um erro médio absoluto de 0,0302, indicando boa preservação das dependências lineares entre os atributos. A pontuação de detecção foi de 0,65, valor que aponta para um bom grau de realismo nos dados gerados.

As comparações visuais reforçam essas evidências. As distribuições de variáveis categóricas, como tipo de produto, linha de produção, turno e resultado final, mostraram alta similaridade entre dados reais e sintéticos. Para variáveis numéricas como SupplierQuality (Figura 1), os gráficos de densidade revelaram padrões compatíveis, ainda que com alguma suavização esperada nesse tipo de abordagem.

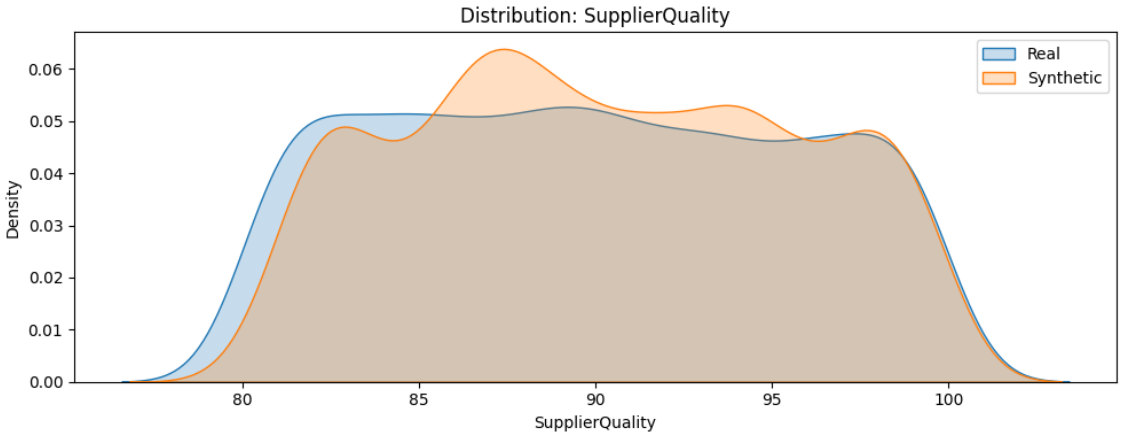

Figura 1 - Densidade de SupplierQuality real vs sintético com curvas similares

A capacidade de gerar dados sintéticos com esse nível de qualidade é relevante em cenários industriais, nos quais a escassez de dados ou restrições de compartilhamento dificultam análises mais robustas. O uso do CTGAN permite ampliar o volume de dados de forma segura, mantendo representatividade estatística e contribuindo para o desenvolvimento de modelos preditivos mais eficazes.

Ainda assim, é necessário reconhecer que a geração de dados sintéticos não substitui completamente os dados reais e que ajustes no modelo podem ser necessários para otimizar a qualidade dos dados produzidos em diferentes contextos. Estudos futuros deverão explorar o uso dos dados gerados para alimentar modelos de machine learning aplicados ao planejamento e controle da produção, avaliando o impacto direto na tomada de decisões.

CONCLUSÕES

Este trabalho demonstrou a viabilidade do uso do modelo CTGAN para geração de dados sintéticos a partir de um conjunto real do setor industrial, evidenciando que a técnica pode produzir amostras com alta fidelidade estatística e distribuição semelhante à dos dados originais. A aplicação dessa abordagem representa uma solução promissora para enfrentar a escassez de dados digitais em ambientes de planejamento e controle da produção, possibilitando a ampliação do volume de dados para o treinamento de modelos preditivos.

Os resultados indicam que os dados sintéticos podem contribuir para a melhoria da análise de desempenho e para a otimização dos processos industriais. Contudo,

destaca-se que o desenvolvimento ainda está em andamento, com etapas futuras previstas para testar a geração de dados sintéticos diretamente em contextos reais de produção e para integrar essas informações em modelos de machine learning para previsão e apoio à tomada de decisão. Assim, a geração de dados sintéticos por meio do CTGAN oferece um caminho inovador para suportar a transformação digital nas indústrias.

REFERÊNCIAS

MIT TECHNOLOGY REVIEW. Bringing Breakthrough Data Intelligence to Industries, 2022. Disponível em: https://www.technologyreview.com/2024/01/09/1085768/bringing-breakthrough-data-intelligence-to-industries/. Acesso em: 25 jul. 2025.

SDMETRICS DEVELOPERS. SDMetrics Documentation. Disponível em: https://docs.sdv.dev/sdmetrics. Acesso em: 5 abr. 2025.