

HUB DE INTELIGÊNCIA ARTIFICIAL  
HUB DE IA / SENAI / FIEP

BARBARA NICOLE ROSA

KATIELE VALÉRIA PEREIRA BRITO

MYLENA HORTZ RIBAS

**MODELOS DE PREVISÃO DOS DADOS DE**

**VENDAS DO SENAI: PESSOA FÍSICA E PESSOA**

**JURÍDICA**

SPRINT 2 – SENAI

LONDRINA – PR

2024

BARBARA NICOLE ROSA

KATIELE VALÉRIA PEREIRA BRITO

MYLENA HORTZ RIBAS

**MODELOS DE PREVISÃO DOS DADOS DE**

**VENDAS DO SENAI: PESSOA FÍSICA E PESSOA**

**JURÍDICA**

SPRINT 1 – SENAI

Relatório para a documentação das Provas de Conceito (POC) para as Empresas/Instituições parceiras do Hub de Inteligência Artificial PR, visando à correta documentação dos conceitos abordados durante a fase de desenvolvimento dos projetos propostos pelos patrocinadores.

LONDRINA – PR

2024

Sumário

[1. Introdução 4](#__RefHeading___Toc1800_657172931)

[1.1. Objetivo 4](#__RefHeading___Toc1802_657172931)

[1.2. Escopo 4](#__RefHeading___Toc1804_657172931)

[1.3. Contra Escopo 5](#__RefHeading___Toc1806_657172931)

[1.4. Riscos 5](#__RefHeading___Toc1808_657172931)

[1.5. Metas Do Projeto 5](#__RefHeading___Toc1810_657172931)

[1.5.1. Entregas Sprint 2 5](#__RefHeading___Toc1812_657172931)

[2. Resultados Alcançados 6](#__RefHeading___Toc1814_657172931)

[2.1. Jupyter Notebook 6](#__RefHeading___Toc1816_657172931)

[2.2. Análises usando Clusterização com a utilização do K-Means, Hierarchical e DBSCAN 6](#__RefHeading___Toc1818_657172931)

[2.2.1. Pessoa Física 6](#__RefHeading___Toc1820_657172931)

[2.3. Dashboard Streamlit 9](#__RefHeading___Toc1822_657172931)

[2.4. Análise e Avaliação de Algoritmos Preditivos 12](#__RefHeading___Toc1824_657172931)

[2.5. Algoritmos 14](#__RefHeading___Toc1826_657172931)

[2.5.1. K-Means 14](#__RefHeading___Toc1828_657172931)

[2.5.2. Hierarchical Clustering 14](#__RefHeading___Toc1830_657172931)

[2.5.3. DBSCAN 15](#__RefHeading___Toc1832_657172931)

[3. Ações de Continuidade 16](#__RefHeading___Toc1834_657172931)

[Referências 16](#__RefHeading___Toc1836_657172931)

# Introdução

O Serviço Nacional de Aprendizagem Industrial (Senai), como parte integrante do Sistema Fiep, desempenha um papel crucial no desenvolvimento industrial e educacional, com mais de 50 unidades distribuídas pelo Paraná e Institutos de Tecnologia em diversas áreas, como por exemplo de Celulose e Papel, Engenharia de Estruturas, Eletroquímica, entre outros.

O Senai também é reconhecido por ser um pilar educacional. Oferece uma ampla gama de cursos, desde iniciação até pós-graduação, incluindo opções presenciais e a distância, atendendo de maneira flexível às necessidades específicas dos alunos e da indústria. Esta diversidade educacional representa um aspecto fundamental de sua atuação, refletindo-se na qualidade de suas análises e contribuições para o setor industrial nacional.

Além do seu compromisso com a educação, o Senai também oferece soluções técnicas, serviços de consultoria, e apoia o desenvolvimento da inovação em parceria com a indústria, com intuito de agilizar o processo de produção e diminuição de gastos com recursos.

## Objetivo

Este projeto visa testar e validar técnicas de clusterização, além de utilizar classificadores usando Machine Learning para agrupar clientes com características semelhantes nos dados fornecidos pelo SENAI. Com o propósito de aprimorar estratégias de marketing direcionadas e personalizadas, assim como identificar tendências de compra específicas para cada segmento de clientes. Essa abordagem busca aumentar a eficiência das campanhas de vendas e aprimorar o atendimento ao cliente.

## Escopo

O escopo deste projeto abrange a aplicação de técnicas de clusterização, o uso de classificadores usando Machine Learning e regras de associação nos dados relacionados às vendas, tanto de pessoas físicas quanto jurídicas, fornecidos pela empresa SENAI. Além disso, será desenvolvido um dashboard utilizando o Streamlit para visualização das análises realizadas.

## Contra Escopo

Não faz parte do escopo deste projeto o desenvolvimento ou entrega de API, bem como o desenvolvimento de uma aplicação web.

## Riscos

Foi identificado como principal risco a possibilidade de insuficiência ou baixa qualidade dos dados fornecidos, incluindo dados incompletos, outliers e erros de anotação. Tais problemas podem resultar em erros de interpretação e imprecisões, comprometendo não apenas a análise imediata, mas também podendo afetar a criação de modelos incapazes de convergir para o resultado desejado.

## Metas Do Projeto

Para a Sprint 2, as metas incluem a aplicação de técnicas de clusterização, o uso de classificadores utilizando Machine Learning e regras de associação com a finalidade de agrupar clientes semelhantes, além da criação de um modelo capaz de recomendar produtos.

## Entregas Sprint 2

Para a entrega da *Sprint 2* foram definidos os seguintes arquivos e documentos:

* *Jupyter Notebook:* com os códigos da clusterização e classificação utilizando Machine Learning e regras de associação nos dados fornecidos, assim como os códigos utilizados para gerar os gráficos utilizados na análise.
* *Dashboard Streamlit:* desenvolvido para visualização iterativa das análises realizadas sem a necessidade de executar os códigos do *Jupyter Notebook*.
* README em *Markdown:* Documento com um tutorial de como utilizar o ambiente virtual para executar o *Jupyter Notebook* e *Dashboard*, incluindo todas as bibliotecas necessárias.
* Relatório final: Documento sobre o desenvolvimento da Sprint 2 e orientações para próximos passos.

# Resultados Alcançados

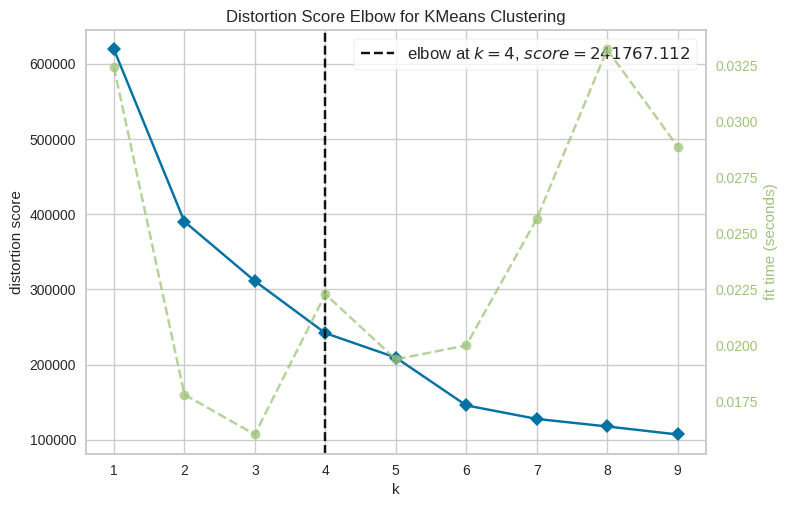
Nesta seção será apresentado os resultados obtidos na *Sprint 2,* a partir de cada uma das metas definidas previamente. Para execução dos entregáveis é necessária a configuração de um ambiente conda. O ambiente conda utilizado foi exportado com todas as bibliotecas necessárias e um tutorial de instalação e uso será disponibilizado juntamente com as entregas no arquivo ‘README.md’. Uma vez instalado na máquina a ser utilizada, é possível executar tanto os arquivos jupyter-notebook quanto o dashboard Streamlit.

## Jupyter Notebook

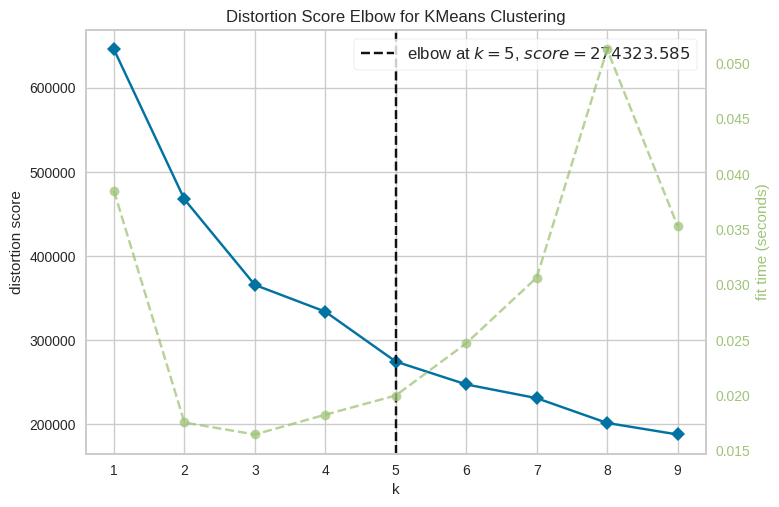
Utilizamos os arquivos já tratados anteriormente, PF\_tratados.csv e PJ\_tratados.csv, para iniciar a aplicação das técnicas de clusterização, o uso de classificadores utilizando Aprendizado de Máquina e regras de associação.

## Análises usando Clusterização com a utilização do K-Means, Hierarchical e DBSCAN

As análises foram iniciadas nos arquivos ‘PF\_tratados.csv’ e ‘PJ\_tratados.csv’, individualmente. Para efetuar a aplicação dos classificadores nos dados, primeiramente realizamos um pré-processamento nos dados, realizando uma transformação nos dados categóricos, excluimos colunas irrelevantes para a realização do método, e por fim fizemos a normalização dos dados.

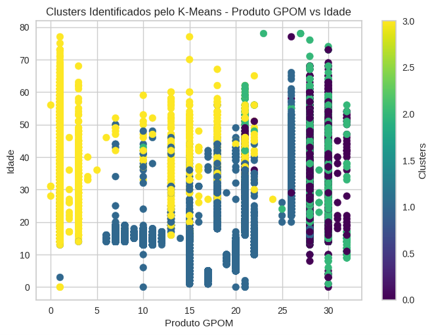
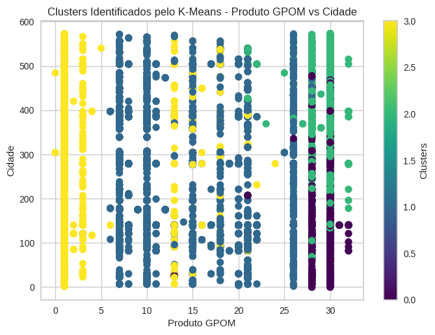
Primeiramente realizamos a aplicação da clusterização utilizando o K-Means, através do método do cotovelo, para determinar o número ótimo de clusters:

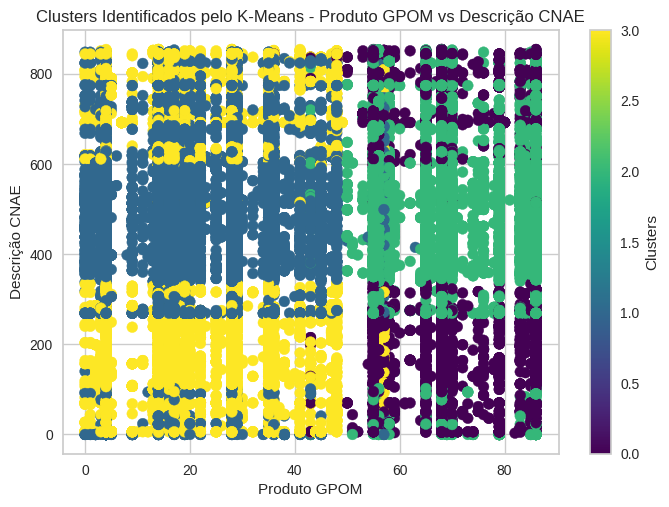
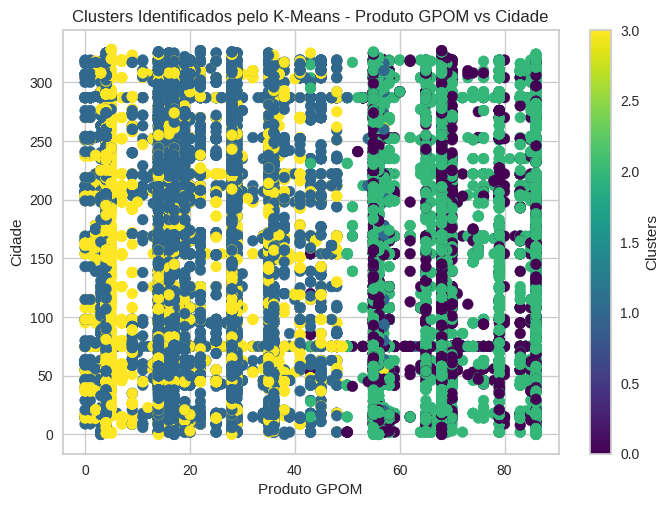
*Figura 1: Método do cotovelo para Pessoa Física, k ótimo k = 4.*



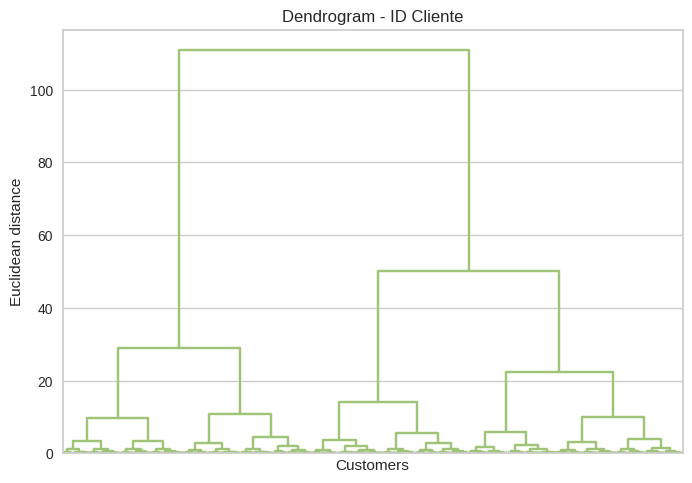
***Figura 2****: Método do cotovelo para Pessoa Jurídica, k ótimo k = 5.*

Logo após, utilizamos uma função para plotar o gráfico de dispersão para todas as colunas possíveis, com o intuito de encontrar os clusters identificados. Cada cluster é representado por uma cor diferente no gráfico, o que facilita a visualização dos grupos formados pela clusterização.

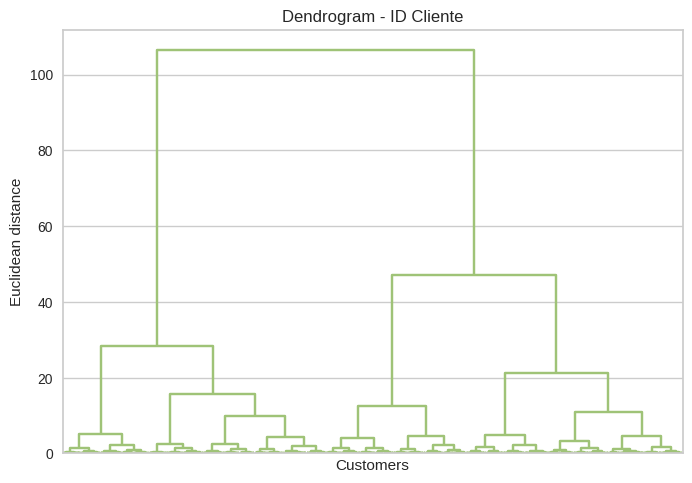
***Figura 3****: Gráfico de dispersão usando K-Means para Pessoa Física.*

***Figura 4****: Gráfico de dispersão usando K-Means para Pessoa Jurídica.*

Em seguida realizamos a aplicação do Hierarchical, usando o dendrograma para visualizar como os dados estão agrupados em diferentes níveis de similaridade. A principal função do dendrograma (Everitt et. al, 2011) é fornecer uma representação visual da relação de proximidade entre os pontos de dados. Cada ponto de dados é representado por uma linha no dendrograma e a proximidade entre esses pontos é representada pela altura das linhas e pelos agrupamentos de linhas próximas.

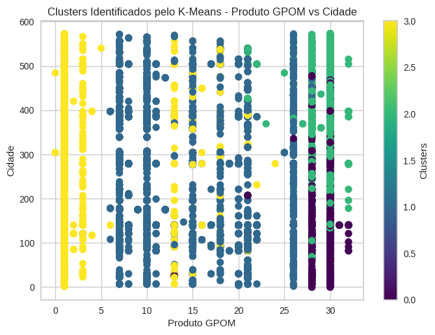
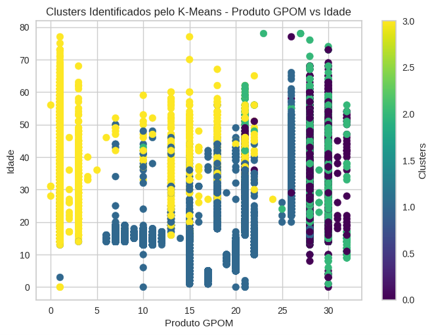


***Figura 5****: Dendrograma do Hierarchical para Pessoa FÍsica.*

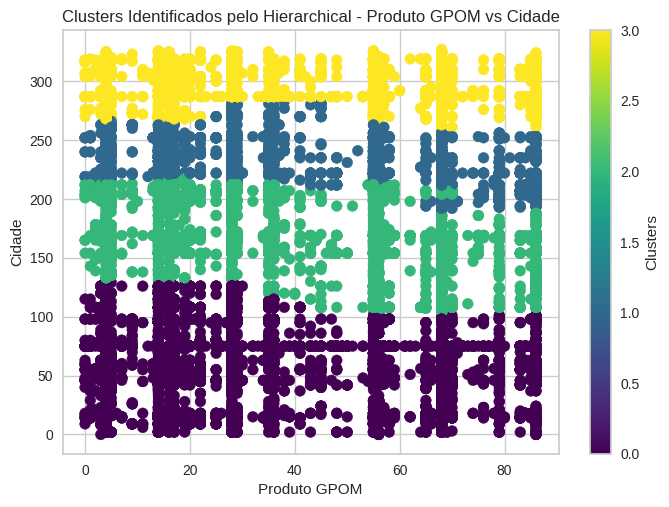
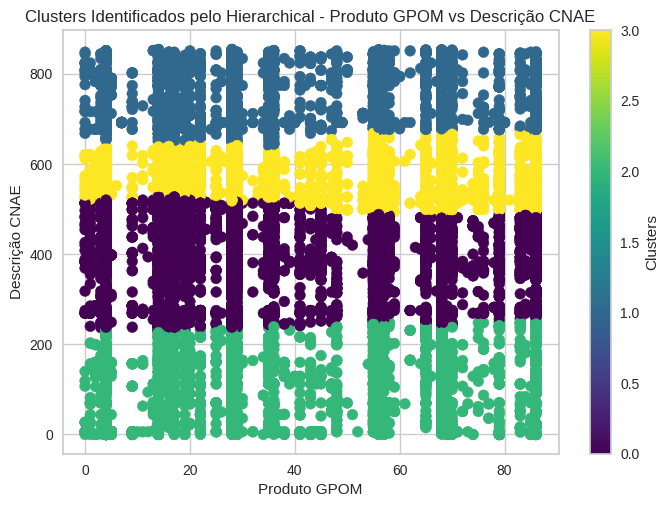


***Figura 6****: Dendrograma do Hierarchical para Pessoa Jurídica.*

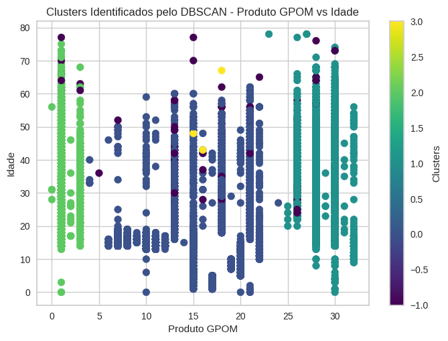
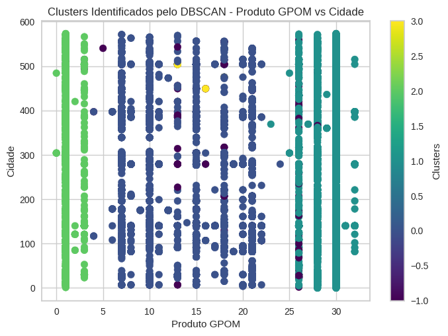
Em seguida, utilizamos o Agglomerative Clustering é um método comum de clusterização hierárquica utilizado para identificar padrões e estruturas em conjuntos de dados. Sua importância reside na capacidade de agrupar os pontos de dados com base na similaridade entre eles. O algoritmo inicia tratando cada ponto como um cluster individual e, em seguida, une iterativamente os clusters mais semelhantes, formando uma estrutura hierárquica. Essa abordagem permite uma compreensão mais profunda das relações entre os dados, facilitando a identificação de grupos naturais e a visualização da estrutura dos dados de maneira intuitiva (Müllner, 2011).

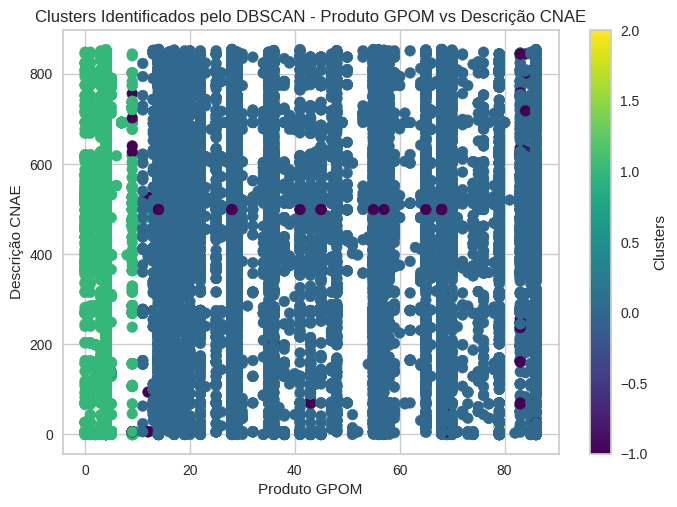
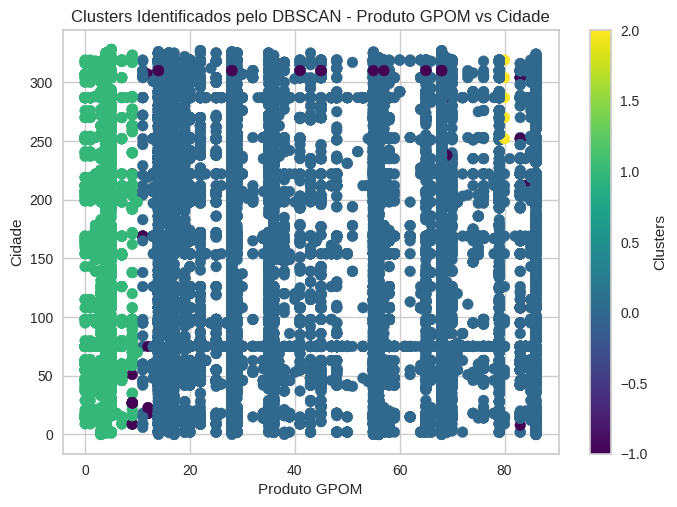


***Figura 7****: Gráfico de dispersão do Hierarchical para Pessoa Física.*

***Figura 8****: Gráfico de dispersão do Hierarchical para Pessoa Jurídica.*

Por fim, utilizamos o algorítimo DBSCAN (Density-Based Spatial Clustering of Applications with Noise) , o DBSCAN é um algoritmo de clustering amplamente utilizado em aprendizado de máquina e análise de dados para identificar clusters de alta densidade em conjuntos de dados com ruído. A principal função do DBSCAN é agrupar os pontos de dados com base na densidade local, o que o torna particularmente eficaz na identificação de clusters de formas arbitrárias. A principal vantagem do DBSCAN é sua capacidade de identificar clusters de forma automática, sem a necessidade de especificar a quantidade de clusters a priori (ESTER, Martin et al., 1996).

***Figura 9****: Gráfico de dispersão do DBSCAN para Pessoa Física.*

***Figura 10****: Gráfico de dispersão do DBSCAN para Pessoa Jurídica.*

## Conclusões

Ao utilizar os algoritmos de clusterização, observou-se que estes não segmentam os clientes com base em outras características disponíveis no banco de dados, deixando de fornecer sugestões personalizadas de produtos que poderiam interessar aos clientes. Isso inclui informações demográficas, idade, genero ou qualquer outro dado disponível que possa revelar padrões de agrupamento. A ausência dessa personalização e segmentação mais refinada limita a capacidade de fornecer recomendações precisas e estratégicas, dificultando a identificação de oportunidades de negócio e ações estratégicas baseadas nos padrões identificados. Acreditamos que isso ocorreu devido ao fato de as técnicas de clusterização utilizarem distâncias entre os pontos. Ao converter os dados categóricos em numéricos, essa abordagem pode ter gerado agrupamentos que não refletem as relações reais entre as variáveis.

* 1. Análises usando Classificação usando Decision Tree, Random Forest e XGBOOST

Vamos utilizar agora análises com classificação utilizando Decision Tree, Random Forest e XGBoost porque a análise utilizando clusterização não foi satisfatória. Apesar dos esforços em identificar padrões e estruturas nos dados por meio da clusterização, não conseguimos segmentar os clientes de forma eficaz com base em outras características disponíveis no conjunto de dados.

Para efetuar a aplicação dos classificadores, primeiramente criamos colunas Targets com os produtos mais vendidos de cada seguimento (Pessoa Física e Pessoa Jurídica), onde 1 representa que o produto foi vendido para o cliente e 0 representa que o produto não vendido, conforme ilustrado na Figura 6. Logo após realização a transformação nos dados categóricos e a normalização.



***Figura 11****: Demonstração da criação de Targets.*

## Métricas Utilizadas

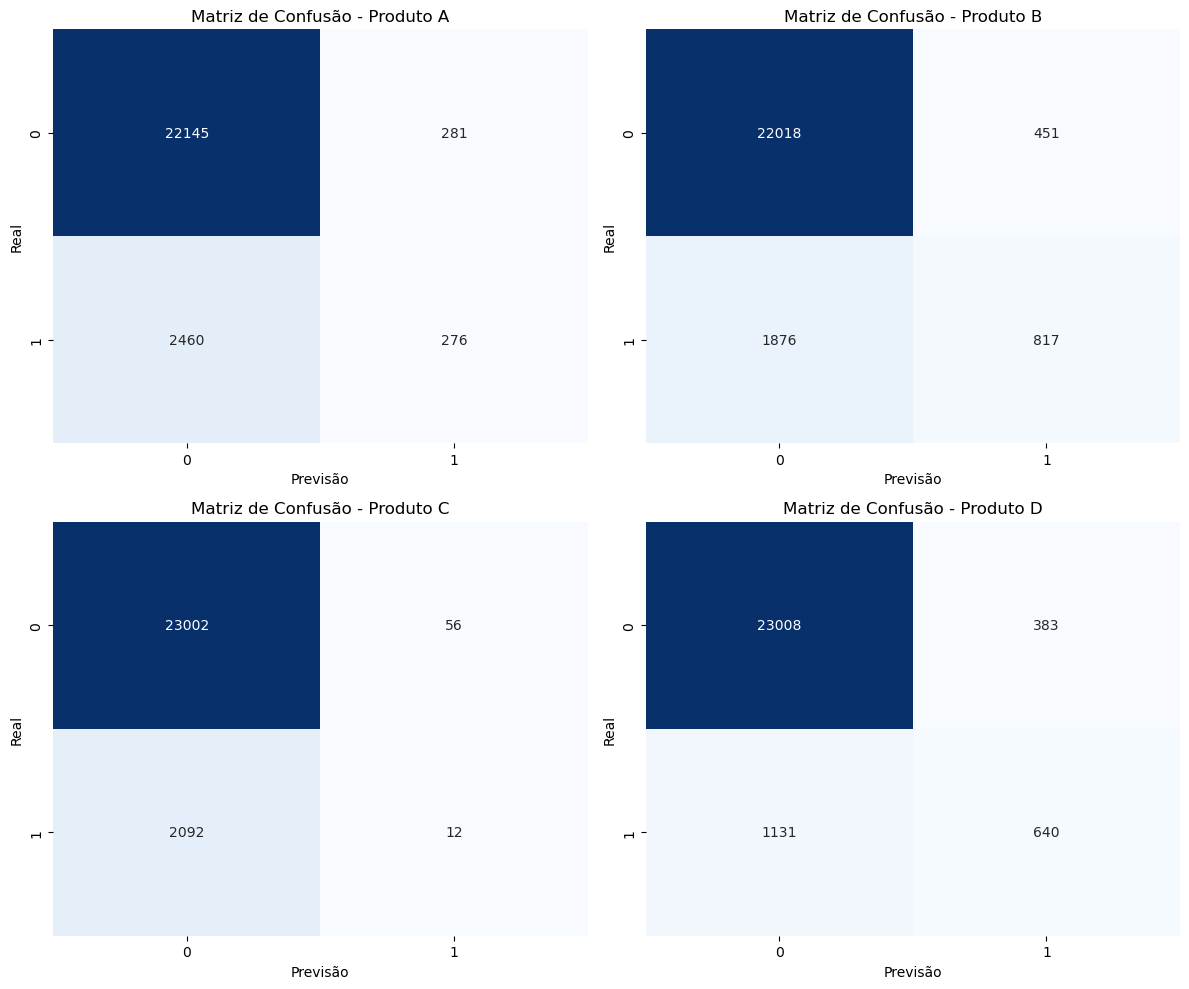
Métricas de avaliação são medidas quantitativas utilizadas para avaliar o desempenho de um modelo de aprendizado de máquina (Dalianis, 2018). São essenciais para compreender como o modelo se comporta em relação aos dados de teste ou validação, auxiliando na determinação se o modelo atende aos requisitos e objetivos estabelecidos (HOSSIN & SULAIMAN, 2015). Dentre as diversas métricas de avaliação disponíveis, as mais comuns incluem acurácia, precisão, recall e F1 Score, e utilizamos estas metricas para os classificadores utilizados.

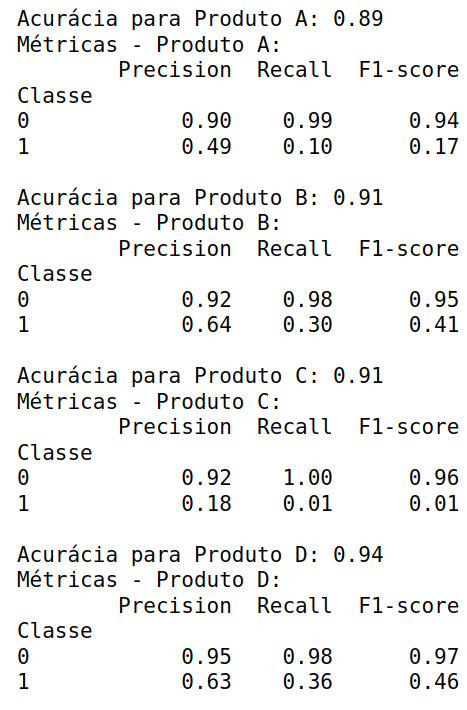
* **Acurácia:** é a medida mais intuitiva de avaliação. Ela representa a proporção de previsões corretas (tanto positivas, quanto negativas) em relação ao total de previsões feitas. Em outras palavras, é a porcentagem de vezes que o modelo acertou.
* **Precisão:** foca na qualidade das previsões positivas do modelo. É a proporção de previsões corretamente positivas (verdadeiros positivos) em relação a todas as previsões positivas (verdadeiros positivos mais falsos positivos).
* **Recall:** mede a capacidade do modelo de detectar todas as instâncias positivas relevantes. É a proporção de verdadeiros positivos em relação ao total de positivos reais (verdadeiros positivos mais falsos negativos). Esta métrica é crucial quando é importante capturar todos os casos positivos possíveis.
* **F1 Score:** é uma métrica que combina precisão e recall em um único número, oferecendo um balanço entre eles.

## Aplicando o classificador Decision Tree

O classificador Decision Tree é um modelo de aprendizado de máquina utilizado para classificação e regressão. Ele constrói uma estrutura de árvore onde cada nó representa uma decisão baseada nos atributos dos dados. A construção da árvore envolve a divisão recursiva dos dados para maximizar a pureza das classes em cada nó folha. Na prática, a árvore faz perguntas sobre os dados, começando com uma característica específica e dividindo os dados em subconjuntos menores com base nas respostas. Esse processo continua até que a árvore chegue a uma decisão final, classificando os dados em categorias distintas (KINGSFORD & SALZBERG, 2008).

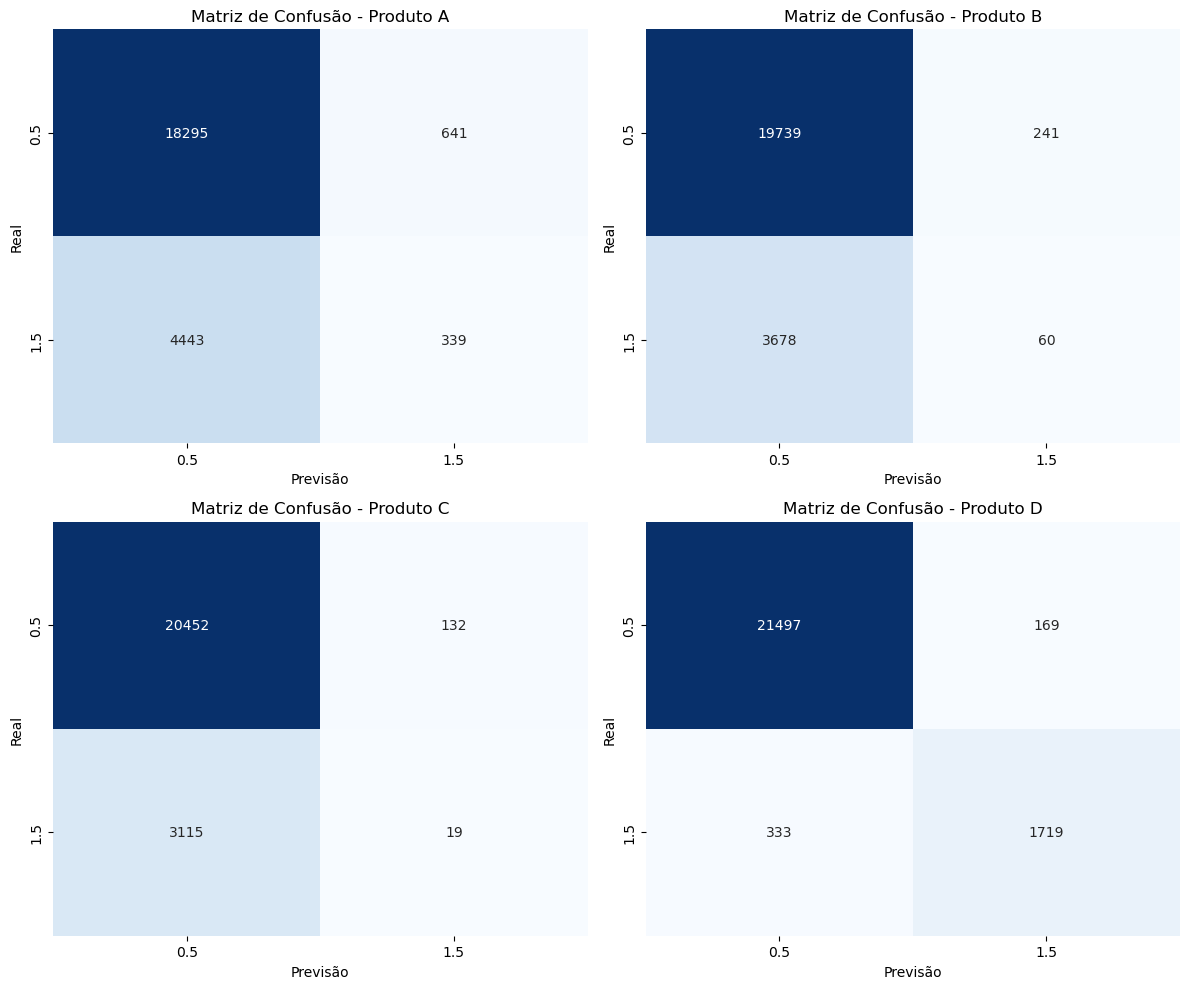
Ao aplicar o Decision Tree, obtivemos a seguinte matriz de confusão e métricas, para quatro produtos distintos dos dados de Pessoa Física:

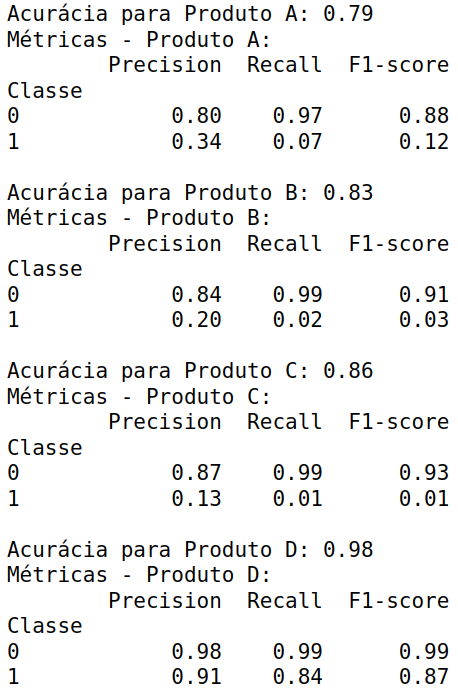




***Figura 12****: Matriz de confusão e Métricas para Pessoa Fisica.*

Para Pessoa Jurídica obtivemos os seguintes resultados:



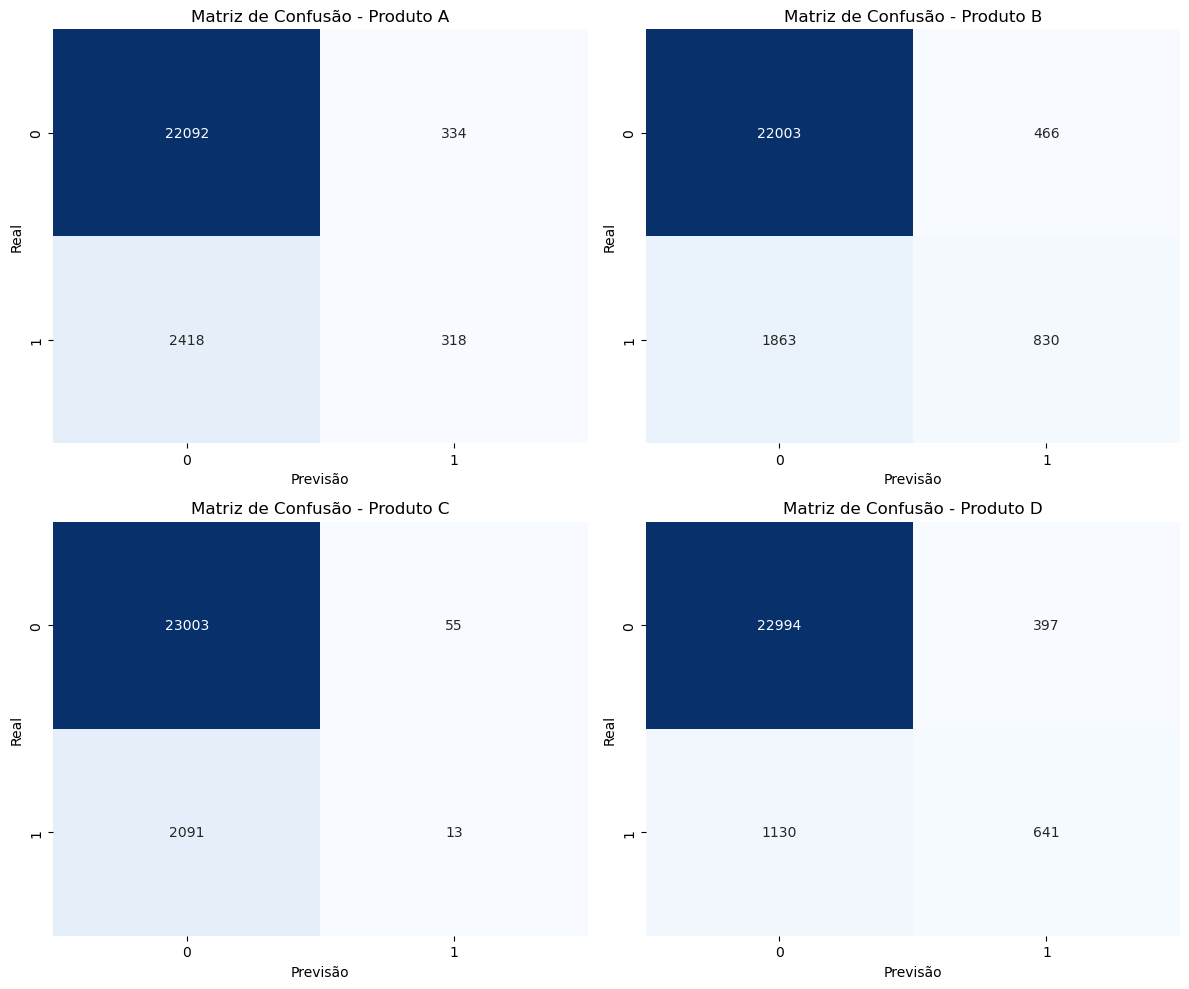


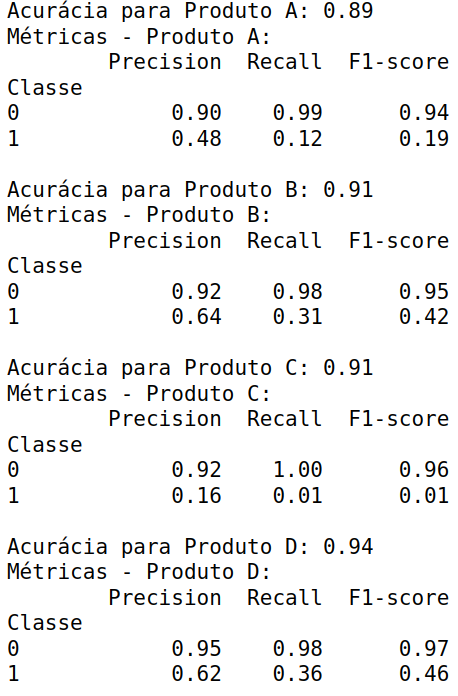
***Figura 13****: Matriz de confusão e Métricas para Pessoa Jurídica.*

## Aplicando o classificador Random Forest

O Random Forest é um modelo de aprendizado de máquina que opera criando uma "floresta" de múltiplas árvores de decisão, cada uma gerada a partir de uma amostra aleatória dos dados. Ao contrário de uma única Árvore de Decisão, o Random Forest combina as previsões de várias árvores para chegar a uma classificação final. Esse processo é conhecido como "ensemble learning". A ideia por trás do Random Forest é que, ao combinar as previsões de várias árvores, cada uma considerando diferentes subconjuntos dos dados e diferentes características, o modelo pode capturar uma representação mais complexa e abrangente dos dados. Isso resulta em um modelo robusto que é eficaz em lidar com uma variedade de conjuntos de dados e tem boa capacidade de generalização (MÜLLER & GUIDO, 2017).

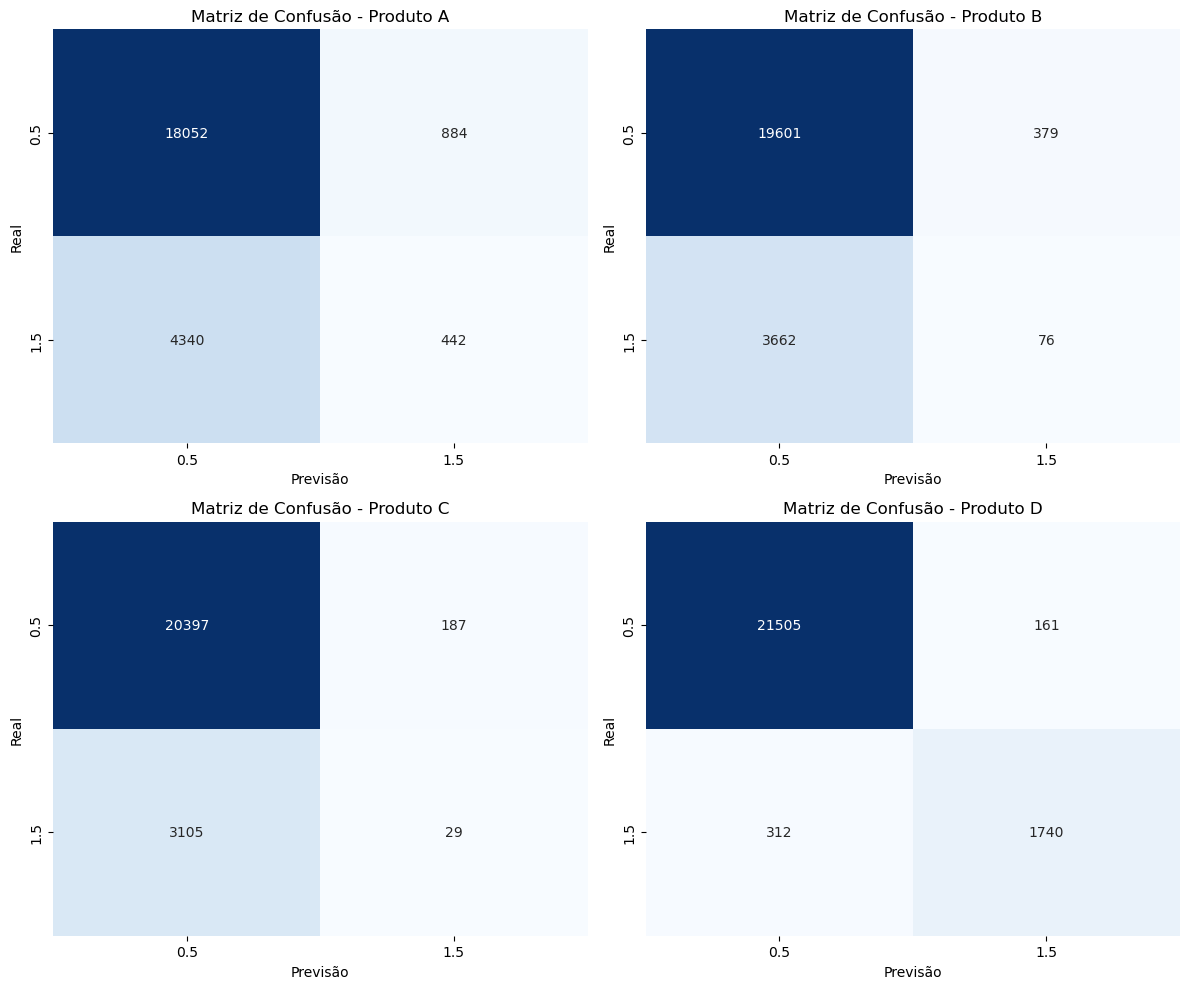
Ao aplicar o Random Forest, obtivemos a seguinte matriz de confusão e métricas, para quatro produtos distintos dos dados de Pessoa Física:

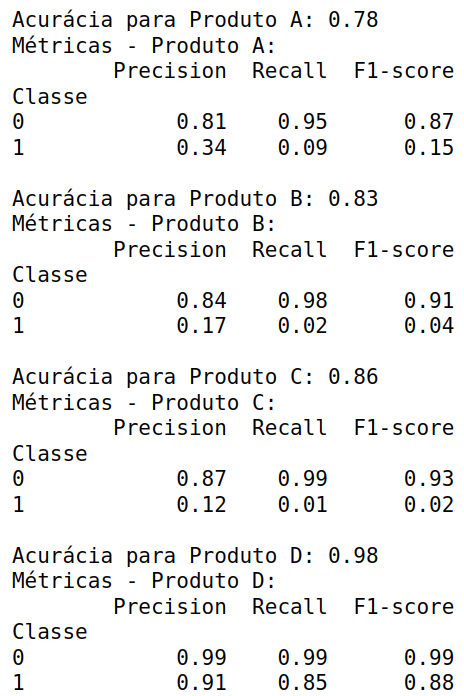




***Figura 14****: Matriz de confusão e Métricas utilizando Random Forest para Pessoa Física.*

Para Pessoa Jurídica obtivemos os seguintes resultados:



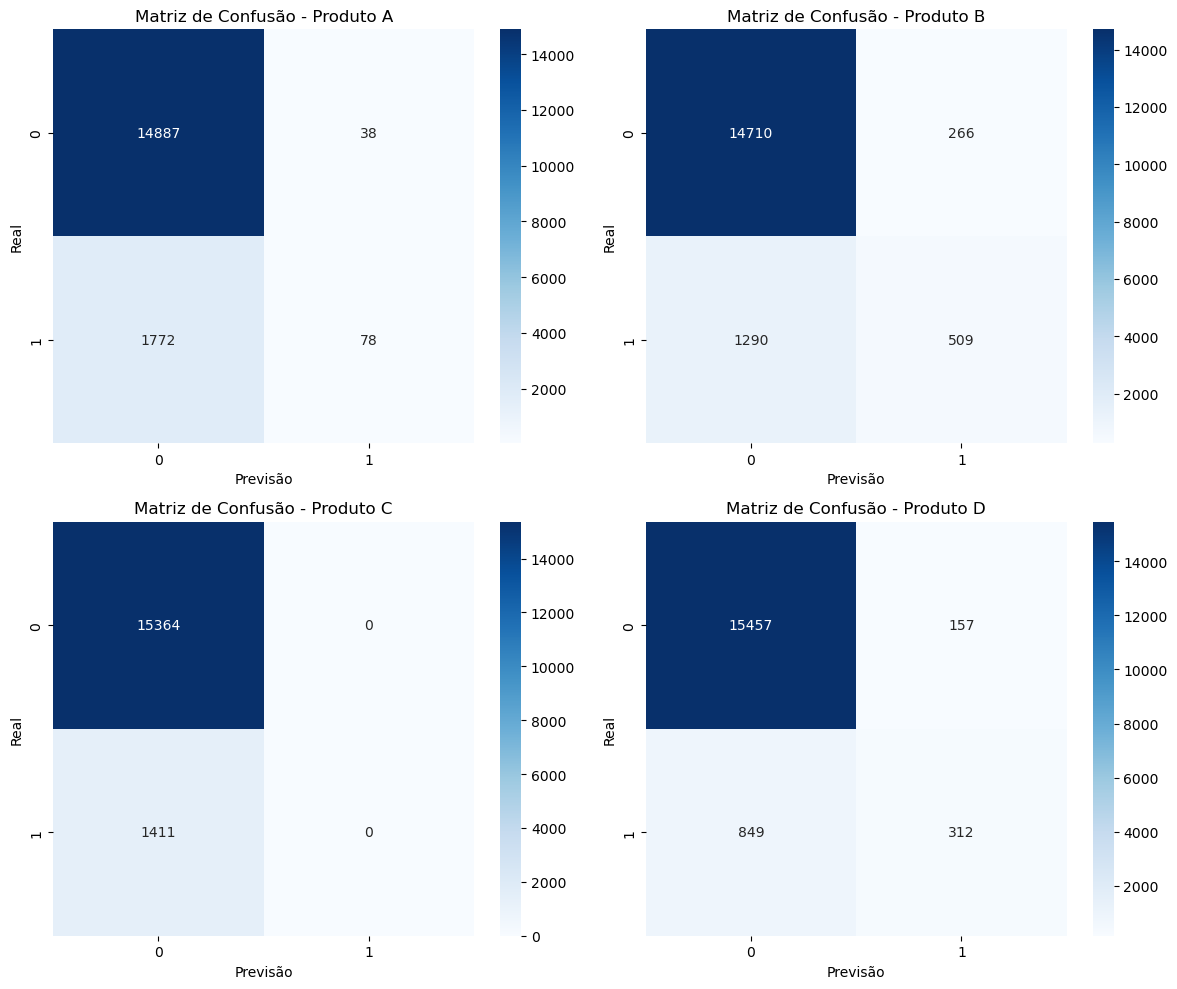


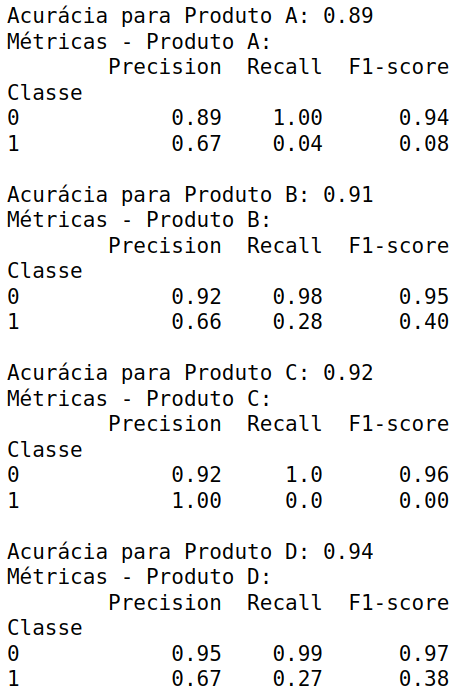
***Figura 15****: Matriz de confusão e Métricas utilizando Random Forest para Pessoa Jurídica.*

## Aplicando o classificador XGBOOST

O XGBoost é uma implementação avançada de aprendizado de máquina, conhecida por sua eficiência e desempenho. Ele opera construindo uma série de árvores de decisão de maneira sequencial, onde cada árvore subsequente tenta corrigir os erros cometidos pelas árvores anteriores. Isso é alcançado dando mais peso aos casos mal previstos nas etapas anteriores, permitindo que o modelo se ajuste e melhore continuamente. No final do processo, as previsões de todas as árvores são combinadas para produzir uma previsão final (CHEN & GUESTRIN, 2016).

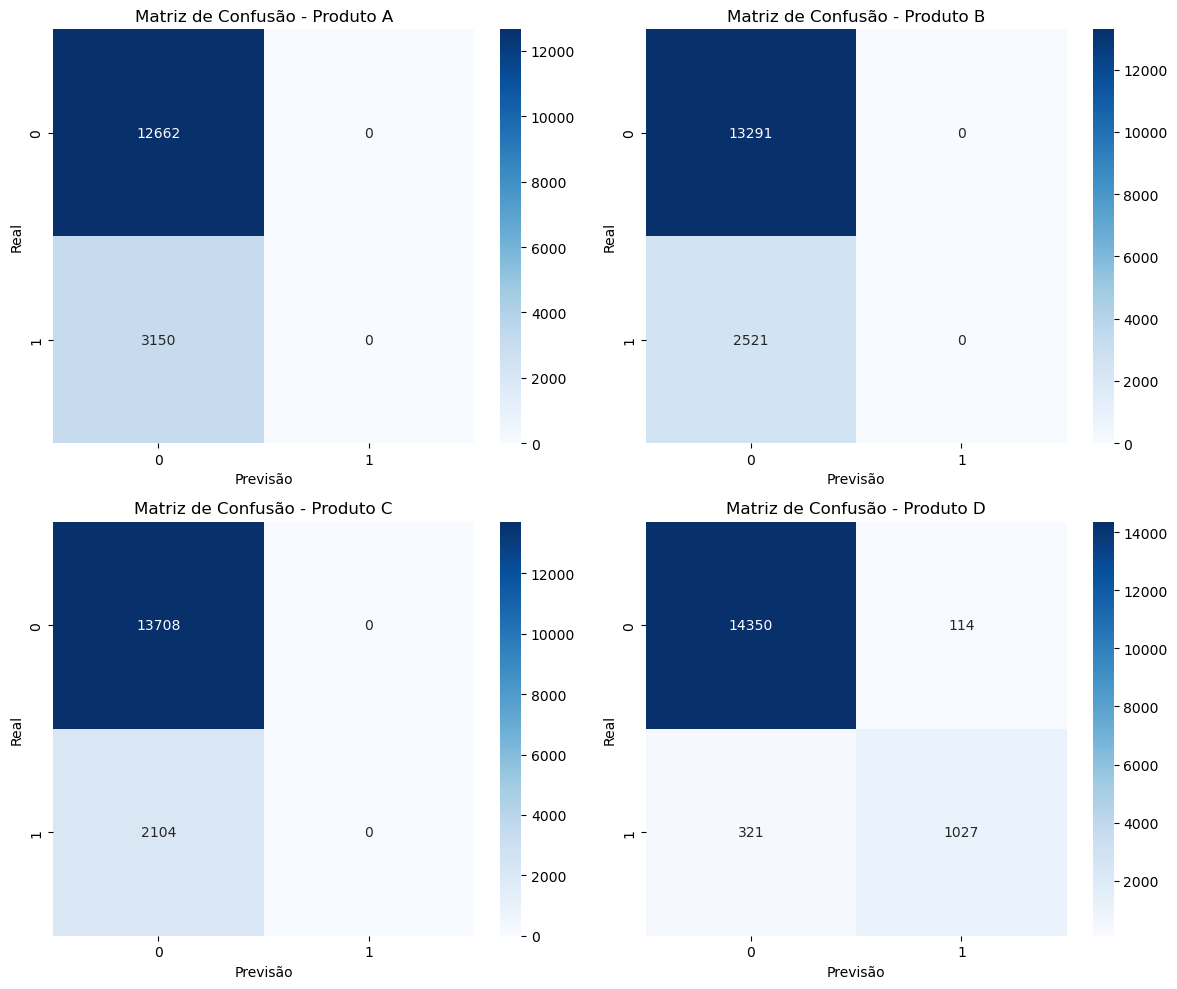
Ao aplicar o XGBOOST, obtivemos a seguinte matriz de confusão e métricas, para quatro produtos distintos dos dados de Pessoa Física:

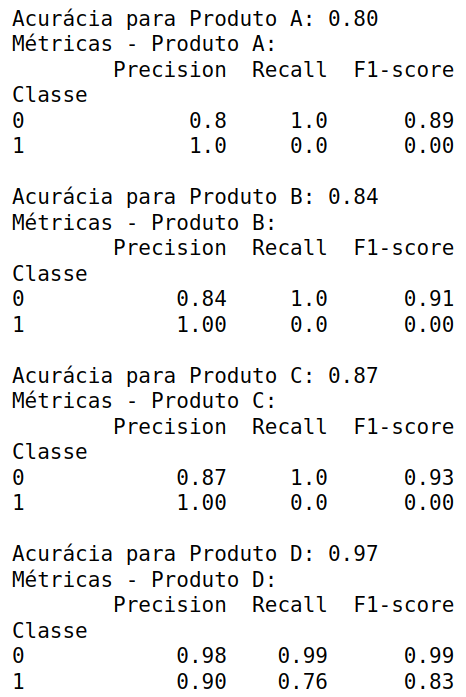




***Figura 16****: Matriz de confusão e Métricas utilizando XGBOOST para Pessoa Física.*

Para Pessoa Jurídica obtivemos os seguintes resultados:





***Figura 17****: Matriz de confusão e Métricas utilizando XGBOOST para Pessoa Física.*

## Conclusões

Com base nos resultados obtidos utilizando os classificadores Decision Tree, Random Forest e XGBoost para os produtos avaliados, observamos que a acurácia, embora geralmente alta, não é uma métrica suficiente para avaliar a eficácia dos modelos. Apesar da alta acurácia, outras métricas importantes, como precisão, recall e F1-score, revelam a incapacidade dos modelos em realizar previsões precisas para todas as classes. Especificamente, notamos que em muitos casos, as métricas para a classe minoritária (Classe 1) são baixas, indicando um desempenho insatisfatório na identificação dessa classe. Isso sugere que os modelos não estão generalizando bem o suficiente para capturar a complexidade dos dados e estão tendo dificuldades em identificar corretamente os casos positivos. Portanto, concluímos que é crucial considerar não apenas a acurácia, mas também outras métricas de avaliação, como precisão, recall e F1-score, para obter uma compreensão mais completa do desempenho dos modelos.

## Análise usando Regra de Associação

## Dashboard Streamlit

Como uma das entregas durante a segunda semana de sprint foi desenvolvido um dashboard no formato Streamlit. O Streamlit é uma biblioteca em python que permite a criação de web apps customizados, e foi escolhido pelo grupo por criar aplicativos intuitivos e de alta interatividade. O código construído no notebook foi criado visando a utilização em um dashboard, desta maneira as funções foram implementadas ao formato Streamlit sem grandes dificuldades.



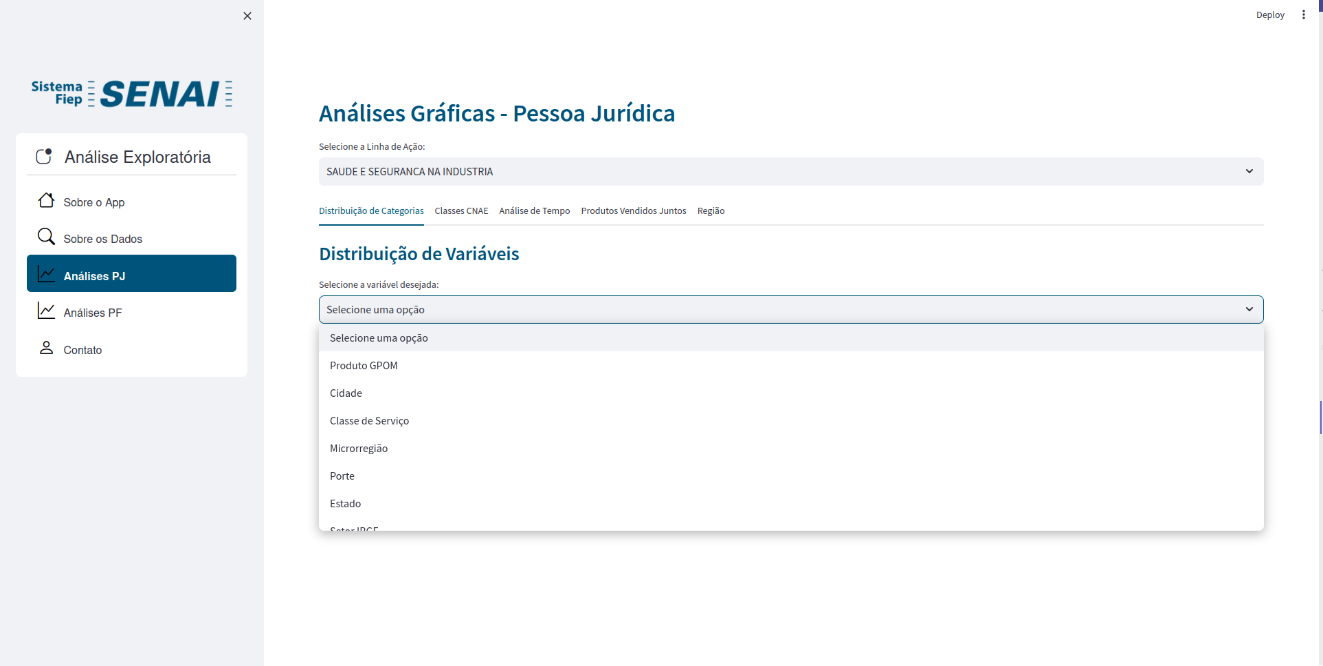
*Figura 18: Página inicial do dashboard desenvolvido.*

Na página inicial temos uma breve introdução e explicação das ferramentas presentes no dashboard. Nesta página, a caixa no canto inferior esquerdo permite o upload de novas bases de dados ao dashboard, desta maneira, uma vez adicionado o novo dataset as análises serão realizadas utilizando os novos dados inseridos. É importante que para as análises serem executadas corretamente o dataset inserido deve estar no mesmo formato do dataset disponibilizado inicialmente, caso contrário o utilizador deverá alterar o código fonte do dashboard pelo arquivo’ dashboard.py’.   
 Na página “Sobre os Dados” é possível visualizar os dados as 10 primeiras linhas dos dados que estão sendo trabalhados, assim como visualizar as categorias presentes em cada uma das colunas.



*Figura 19: Página de visualização dos dados.*

Na página “Análises PJ”, no topo é possível selecionar a Linha de Ação, que inicia já seleciona a de maior relevância. As análises estão divididas em abas: Distribuição de Categorias, Classes CNAE, Análise de Tempo, Produtos Vendidos Juntos e Região. Possuindo diversas visualizações gráficas interativas das variáveis e relações entre elas de acordo com cada análise.



***Figura 20:*** *Página de análises gráficas PJ.*

Seguindo a mesma lógica, no topo da página “Análises PF” é possível selecionar a Linha de Ação, que inicia já seleciona a de maior relevância. As análises estão divididas em abas: Distribuição de Categorias, Análise de Tempo, Idade dos Clientes, Gênero e Vendas, Produtos Vendidos Juntos e Região. Possuindo diversas visualizações gráficas interativas das variáveis e relações entre elas de acordo com cada análise.



***Figura 21:*** *Página de análises gráficas PF.*

# Ações de Continuidade

# Referências

Everitt, B. S., Landau, S., Leese, M., & Stahl, D. (2011). Cluster Analysis. –John Wiley & Sons. *Ltd., New York*, 330.

Müllner, D. (2011). Algoritmos modernos de agrupamento hierárquico e aglomerativo. *Pré-impressão do arXiv arXiv:1109.2378* .

ESTER, Martin et al. Um algoritmo baseado em densidade para descobrir clusters em grandes bancos de dados espaciais com ruído. Em: kdd . 1996. pág. 226-231.

DALIANIS, Hercules; DALIANIS, Hercules. Evaluation metrics and evaluation. Clinical Text Mining: secondary use of electronic patient records, p. 45-53, 2018.

HOSSIN, Mohammad; SULAIMAN, Md Nasir. A review on evaluation metrics for data classification evaluations. International journal of data mining & knowledge management process, v. 5, n. 2, p. 1, 2015.

KINGSFORD, Carl; SALZBERG, Steven L. What are decision trees?. Nature biotechnology, v. 26, n. 9, p. 1011-1013, 2008.

Andreas C. Müller & Sarah Guido . Introduction to machine learning with python. 2017.

CHEN, Tianqi; GUESTRIN, Carlos. Xgboost: A scalable tree boosting system. In: Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining. 2016. p. 785-794.