INSTITUTO TECNOLÓGICO DE AERONÁUTICA

Hermiro da Cruz Pessoa Junior

ANÁLISE DE PERFIS SOCIOECONÔMICOS E PREDIÇÃO DA SATISFAÇÃO DE PASSAGEIROS EM AEROPORTO UTILIZANDO MACHINE LEARNING

Trabalho de Graduação
2025

Curso de Engenharia Civil-Aeronáutica

CDU 656.7

Hermiro da Cruz Pessoa Junior

ANÁLISE DE PERFIS SOCIOECONÔMICOS E PREDIÇÃO DA SATISFAÇÃO DE PASSAGEIROS EM AEROPORTO UTILIZANDO MACHINE LEARNING

Orientador

Prof. Alessandro Vinícius Marques de Oliveira (ITA)

ENGENHARIA CIVIL-AERONÁUTICA

São José dos Campos
Instituto Tecnológico de Aeronáutica

2025

Dados Internacionais de Catalogação-na-Publicação (CIP)
Divisão de Informação e Documentação

Pessoa Junior, Hermiro da Cruz
ANÁLISE DE PERFIS SOCIOECONÔMICOS E PREDIÇÃO DA SATISFAÇÃO DE PASSAGEIROS EM AEROPORTO UTILIZANDO MACHINE LEARNING / Hermiro da Cruz Pessoa Junior.
São José dos Campos, 2025.
36f.

Trabalho de Graduação – Curso de Engenharia Civil-Aeronáutica– Instituto Tecnológico de Aeronáutica, 2025. Orientador: Prof. Alessandro Vinícius Marques de Oliveira.

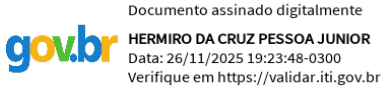
1. . 2. . 3. . I. Instituto Tecnológico de Aeronáutica. II. ANÁLISE DE PERFIS SOCIOECONÔMICOS E PREDIÇÃO DA SATISFAÇÃO DE PASSAGEIROS EM AEROPORTO UTILIZANDO MACHINE LEARNING.

REFERÊNCIA BIBLIOGRÁFICA

PESSOA JUNIOR, Hermiro da Cruz. **ANÁLISE DE PERFIS SOCIOECONÔMICOS E PREDIÇÃO DA SATISFAÇÃO DE PASSAGEIROS EM AEROPORTO UTILIZANDO MACHINE LEARNING**. 2025. 36f. Trabalho de Conclusão de Curso (Graduação) – Instituto Tecnológico de Aeronáutica, São José dos Campos.

CESSÃO DE DIREITOS

NOME DO AUTOR: Hermiro da Cruz Pessoa Junior
TITULO DO TRABALHO: ANÁLISE DE PERFIS SOCIOECONÔMICOS E PREDIÇÃO DA SATISFAÇÃO DE PASSAGEIROS EM AEROPORTO UTILIZANDO MACHINE LEARNING.
TIPO DO TRABALHO/ANO: Trabalho de Conclusão de Curso (Graduação) / 2025



Hermiro da Cruz Pessoa Junior
Rua H8B, Ap. 238
12.228-461 – São José dos Campos–SP

ANÁLISE DE PERFIS SOCIOECONÔMICOS E PREDIÇÃO DA SATISFAÇÃO DE PASSAGEIROS EM AEROPORTO UTILIZANDO MACHINE LEARNING

Essa publicação foi aceita como Relatório Final de Trabalho de Graduação

Documento assinado digitalmente
gov.br
HERMIRO DA CRUZ PESSOA JUNIOR
Data: 26/11/2025 19:24:46-0300
Verifique em https://validar.iti.gov.br

Hermiro da Cruz Pessoa Junior

Autor

Documento assinado digitalmente
gov.br
ALESSANDRO VINICIUS MARQUES DE OLIVEIRA
Data: 26/11/2025 20:28:42-0300
Verifique em https://validar.iti.gov.br

Alessandro Vinícius Marques de Oliveira (ITA)

Orientador

São José dos Campos, 13 de novembro de 2025.

Agradecimentos

Primeiramente, gostaria de agradecer à minha família: meus pais, Hermiro e Nazaré, por terem me criado com tanto carinho e por me transmitirem os valores de dedicação, honestidade e persistência.Gostaria de agradecer também a meus irmãos, Giovanni e Leticia, por sempre serem um ponto de apoio em todos os momentos. Obrigado por me apoiarem em todas as minhas decisões e por celebrarem cada pequena vitória ao meu lado.

Aos meus amigos de H8 pelas risadas e pelos momentos de descontração. Em especial para o Zé, Gabriel, Moreira, Luiz e Eduardo. Cheguei em São José conhecendo pouca gente e saio com amigos para a vida toda.

Além disso, gostaria de agradecer em especial aos amigos que a Civil me proporcionou, em especial Juliana, João, Felipe e Mioni. É certo dizer que sem a ajuda de vocês, essa graduação não teria acontecido.

E, por fim, a todos aqueles que, de alguma forma, contribuíram para a realização deste projeto e para o meu crescimento pessoal e profissional. Obrigado por fazerem parte da minha história.

Resumo

Este trabalho apresenta uma análise aprofundada dos perfis socioeconômicos dos passageiros e o desenvolvimento de um modelo robusto de Machine Learning (ML) para predição de sua satisfação. A pesquisa foi conduzida utilizando uma pesquisa de satisfação de passageiros do Aeroporto Internacional de Guarulhos(GRU) e do Aeroporto Internacional de Brasília(BSB) entre os anos de 2018 e 2021, abrangendo variáveis como gênero, frequência de voo, escolaridade. A metodologia envolveu um pré-processamento de dados, uma clusterização com K-Means, seguida da implementação do algoritmo de Random Forest para ML. Os resultados demonstraram que o gênero, a forma de acesso ao aeroporto e a presença ou não de conexão são fatores importantes para o agrupamento de passageiros aéreos. O modelo de ML final, baseado em Random Forest, atingiu uma precisão de 80.6% (86.1% de F1-Score) na predição da satisfação. Além disso, concluiu-se que o embarque é um básico de qualidade. No entanto, para diferenciar e alcançar alta satisfação acima da média,se observou que aspectos como Locomoção Interna e conforto ambiental se tornam diferenciais.

Abstract

This study presents an in-depth analysis of passenger socioeconomic profiles and the development of a robust Machine Learning (ML) model for predicting their satisfaction. The research utilized a passenger satisfaction survey conducted at Guarulhos International Airport (GRU) and Brasília International Airport (BSB) between 2018 and 2021, covering variables such as gender, flight frequency, and education level.

The methodology involved data pre-processing, followed by K-Means clustering, and the implementation of the Random Forest algorithm for ML prediction. The results demonstrated that gender, means of access to the airport, and the presence of a connection are important factors for grouping air passengers.

The final ML model, based on Random Forest, achieved an accuracy of 80.6% (F1-Score of 86.1%) in predicting satisfaction. Furthermore, it was concluded that the boarding process is a basic quality requirement. However, to differentiate and achieve above-average high satisfaction, aspects such as internal locomotion and environmental comfort become key differentiators.

Lista de Figuras

Lista de Tabelas

Sumário

1 Introdução

A experiência do passageiro em aeroportos é um ponto estratégico para a competitividade e as receitas, especialmente em hubs de alta demanda, como o Aeroporto Internacional de São Paulo/Guarulhos (GRU) e o Aeroporto Internacional de Brasília (BSB). Nesse contexto, os programas de avaliação padronizada, como o *Airport Service Quality* (ASQ) do ACI, consolidaram indicadores que capturam percepções de qualidade e satisfação,servindo como referência para investimentos em infraestrutura e operação. Paralelamente, estudos demonstraram que a satisfação é resultado de múltiplas dimensões do serviço aeroportuário (acessibilidade, conforto, lojas, limpeza, etc.), percebidas distintamente pelas características dos próprios passageiros.(Brasil. Secretaria Nacional de Aviação Civil (SAC), 2024)

Nesse contexto, a segmentação dos passageiros em perfis socioeconômicos é importante para evitar decisões baseadas em médias que ocultam preferências distintas; a identificação de grupos homogêneos permite especificar serviços e corrigi-los com maior precisão. Além disso, é notável a implementação de modelos preditivos que indiquem, com boa interpretabilidade, os principais fatores que contribuem para a satisfação e o que mais pesa em cada perfil e no agregado. Assim, métodos de *machine learning* já se mostram adequados para prever e explicar a satisfação no transporte aéreo, oferecendo um caminho aplicável para decisões baseadas em evidências. Embora modelos de *Machine Learning* frequentemente operem como ”caixas pretas”,o estudo em questão busca averiguar a interpretabilidade dos resultados, a fim de fornecer análises compreensíveis para a gestão aeroportuária.(Almeer; Alfayez, 2022; Şahinbaş, 2022)

Este trabalho procura abordar essa temática utilizando dados reais da pesquisa de satisfação do passageiro, uma iniciativa contínua da Agência Nacional de Aviação Civil (ANAC), focando em GRU e (BSB). Foi aplicada o método de *clusterização* por *K-Means* para identificar perfis socioeconômicos de passageiros,e além disso, foi utilizado o método de aprendizado de máquina por *Random Forest* para prever satisfação e ranquear os fatores mais influentes por perfil e no agregado. O objetivo prático desse trabalho é entender nesses métodos quais as melhores abordagens no tratamento de bases reais para que se tenha um melhor suporte nas decisões de atendimento especializado ao passageiro.

2 Revisão Bibliográfica

2.1 Clusterização com K-Means

O algoritmo K-Means é um método de particionamento amplamente utilizado para minimizar a soma das distâncias quadráticas intragrupo (SSE), tipicamente resolvido por iterações Lloyd (realocação e atualização de centróides). A qualidade da solução depende fortemente da inicialização dos centróides; por isso, a estratégia K-Means++ propõe um *seeding* probabilístico que favorece pontos afastados. Este favorecimento é alcançado ao selecionar os centróides com uma probabilidade proporcional ao quadrado de suas distâncias em relação aos centróides já escolhidos, o que provê garantias teóricas de aproximação e melhora a convergência. Assim, implementações modernas (*e.g.*, `scikit-learn`) disponibilizam K-Means com inicialização K-Means++ como padrão, tornando a prática reprodutível e eficiente, mas a escolha ideal deve sempre refletir as particularidades do domínio probabilístico da aplicação.(Qian; Li; Wang, 2024)

2.1.1 Objetivo do K-Means e minimização da soma intragrupo

Dado um conjunto de dados $\mathcal{X} = \{\mathbf{x}_i\}_{i=1}^{n} \subset \mathbb{R}^d$ e k clusters, o K-Means busca a partição $\{C_1, \ldots, C_k\}$ e centróides $\{\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_k\}$ que minimizam a soma das distâncias quadráticas intragrupo (WCSS/SSE):

$$\mathcal{J}(\{C_k\}, \{\boldsymbol{\mu}_k\}) = \sum_{k=1}^{K} \sum_{\mathbf{x}_i \in C_k} \|\mathbf{x}_i - \boldsymbol{\mu}_k\|_2^2 . \tag{2.1}$$

Passo de atribuição (fixos $\boldsymbol{\mu}_k$): para centróides fixos, a minimização de (2.1) implica atribuir cada ponto ao cluster do centróide mais próximo:

$$\mathbf{x}_i \in C_{k^\star} \quad \text{com} \quad k^\star = \arg\min_{1 \le k \le K} \|\mathbf{x}_i - \boldsymbol{\mu}_k\|_2^2 . \tag{2.2}$$

Passo de atualização (fixos C_k): para partições fixas, a derivada de (2.1) em relação a

$\boldsymbol{\mu}_k$ mostra que o minimizador é a média dos pontos no cluster k:

$$\boldsymbol{\mu}_k = \frac{1}{|C_k|} \sum_{\mathbf{x}_i \in C_k} \mathbf{x}_i. \tag{2.3}$$

O andamento do método de K-Means passa pelo Algoritmo de Lloyd, que consiste em iterar alternadamente (2.2) e (2.3) até convergir.Vale lembrar que a cada iteração $\mathcal{J}$ não aumenta e o algoritmo converge para um mínimo local da (2.1) (Lloyd, 1982).

2.1.2 Escolha do número de clusters (k)

2.1.2.1 Elbow Method (critério do "cotovelo")

O Elbow Method é um método para escolher o número de clusters k em K-Means observando a redução da variabilidade dentro das partições ao aumentar k. A ideia central desse método é traçar uma curva dos pesos (como (2.1)) para cada cluster em função de $k \in \{2, \ldots, K_{\max}\}$. Para valores pequenos de k, cada aumento de k costuma gerar alta diminuição na soma dos pesos e, a partir de certo ponto, a redução torna-se menos expressiva.Assim, escolhe-se k no ponto de inflexão visual (o "cotovelo").

Na prática, o Elbow Method fornece um intervalo plausível de k (em especial quando o "cotovelo" é visualmente nítido. Vale lembrar que o "cotovelo" pode ser subjetivo ou inexistente (curva com tendencia linear ou bases com forte ruído). Nesses casos, é interessante ter o método apenas como triagem, complementando com outras modalidades de escolha do k.

Como complemento, o *Silhouette Score* mede, para cada observação, a coesão (proximidade ao seu cluster) versus separação (distância aos demais), variando de -1 a $+1$; a média global auxilia a comparar soluções com diferentes k e a inspecionar qualidade por cluster . Em dados reais de serviços (como aeroportos), a análise conjunta de Elbow e Silhouette é prática comum: o primeiro dá um "range" plausível de k, enquanto o segundo prioriza configurações com melhor separação/coesão — especialmente quando perfis socioeconômicos têm fronteiras parcialmente sobrepostas.

2.1.2.2 Silhouette Score

O *Silhouette Score* avalia simultaneamente a coesão(proximidade ao seu cluster) de um ponto ao seu próprio cluster e a separação em relação ao cluster vizinho mais próximo. Para cada observação i, define-se: $a(i)$ como a distância média de i aos demais pontos do seu cluster, e $b(i)$ como a menor distância média de i aos pontos de qualquer outro

cluster.O índice para um elemento do cluster é:

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i),\, b(i)\}} \in [-1,\, 1], \tag{2.4}$$

onde:

- Os valores próximos de 1 indicam boa classificação do elemento (boa proximidade ao cluster escolhido e alta distância do outro cluster);
- Os valores próximos de 0 sugerem a proximidade do elemento a uma fronteira entre os respectivos clusters;
- Os valores negativos indicam uma má classificação desse elemento (uma proximidade maior ao outro cluster do que ao escolhido).

O *Silhouette médio*, também chamado de *silhouette score*, é a média de $s(i)$ em toda a amostra (ou por cluster), permitindo comparar soluções com diferentes k e identificar clusters com melhores desempenho. Nisso, vale ressaltar que esse método tende a favorecer clusters aproximadamente esféricos, e para altos valores de k, as diferentes distâncias podem diminuir, o que reduz o poder discriminante.

Na literatura, se encontra em (Şahinbaş, 2022) um silhouette médio de 0.145 para clusterização de passageiros aéreos.

2.2 Algoritmo do Random Forest

2.2.1 Random Forest para Predição de Classificação

O *Random Forest* (RF) é um método de aprendizado de máquina baseado em múltiplas árvores de decisão, treinadas sobre amostras *bootstrap* (amostragem com reposição) e com aleatorização na seleção de atributos(ou *features* - normalmente as colunas escolhidas em uma base de dados). A combinação de *bagging*(treinamento independente dos subconjuntos de árvoes de decisão) com a escolha aleatória de subconjuntos de *features* em cada divisão confere ao RF uma boa acurácia em cenários com relações não-lineares ou alta dimensionalidade dos dados.

Na construção do modelo,para cada árvore $b = 1, \ldots, B$:

- Sorteia-se com reposição uma amostra de tamanho n do conjunto de treino;
- em cada nó, avalia-se um subconjunto aleatório de m atributos ($m <$ número de features) e escolhe-se a partição que maximiza a redução de impureza;

- cresce-se a árvore até um critério de parada (*max_depth*, *min_samples_leaf* etc.)

Na classificação, a predição final é a classe(alvo da predição) com maior número de votos entre as B árvores,Nisso, as probabilidades estimadas podem ser obtidas como a fração de votos por classe.

Nas árvores do RF,o critério de a qualidade de uma partição é medida por uma função de impureza do nó. Para o índice de Gini, a impureza é interpretada como o quão misto é um nó de decisão: é a probabilidade de errar a classe ao rotular aleatoriamente uma amostra segundo as proporções de classes do próprio nó (0 em nó "puro" e maior quanto mais equilibradas as classes). Para isso, se faz conveniente projetar features que permitam cortes "limpos", como as flags binárias e as ordenações.Assim, quanto mais um atributo consegue formar nós com alta pureza (Gini baixo), mais rapidamente ele aparece na árvore, e mais forte em correlação ele tende a ser para explicar os principais fatores de satisfação.

2.2.2 Métricas de avaliação de modelos de classificação

Considere a *confusion matrix* com Verdadeiro Positivo (TP), Falso Positivo (FP), Verdadeiro Negativo (TN) e Falso Negativo (FN). Com base nesses termos:

Acurácia (Accuracy). Proporção de acertos totais. Útil quando as classes estão balanceadas.

$$\text{ACC} = \frac{TP + TN}{TP + FP + TN + FN}. \tag{2.5}$$

Precisão (Precision) Entre as previsões positivas, quantas são realmente positivas (controle de falsos positivos).

$$\text{Precision} = \frac{TP}{TP + FP}. \tag{2.6}$$

Revocação/Sensibilidade (Recall/TPR) Entre as positivas reais, quantas o modelo detecta (controle de falsos negativos).

$$\text{Recall} = \text{TPR} = \frac{TP}{TP + FN}. \tag{2.7}$$

Especificidade (TNR) Entre as negativas reais, quantas o modelo acerta.

$$\text{Specificity} = \text{TNR} = \frac{TN}{TN + FP}. \tag{2.8}$$

F1-Score Média harmônica entre Precisão e Recall; útil com classes desbalanceadas.

$$\text{F1} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}. \tag{2.9}$$

Curva ROC e AUC-ROC ROC plota TPR (eixo y) vs. FPR = $\frac{FP}{FP+TN}$ (eixo x) ao variar o limiar. AUC-ROC é a *área sob a curva* (aproximada por soma de trapézios). Modelos melhores têm AUC mais próxima de 1.

Curva Precisão–Revocação e AUC-PR Plota Precisão (eixo y) vs. Recall (eixo x) ao variar o limiar; em classes raras, é mais informativa que ROC. AUC-PR é a área sob essa curva (regra dos trapézios).(Goldschmidt; Bezerra; Passos, 2015)

Na literatura, encontra-se em (Cunha, 2023), valores acima de 70% para acurácia, precisão e recall para modelos de classificação utilizando Machine Learning em satisfação de passageiros.

3 Metodologia

A metodologia empregada seguiu uma sequência lógica e interligada de três etapas cruciais. Inicialmente, o Pré-processamento de Dados é responsável pela limpeza, transformação e normalização do conjunto de dados, tratando valores ausentes, outliers e inconsistências, minimizando o ruído e otimizando. Em seguida, a Clusterização foi aplicada para identificar grupos naturais ou segmentos dentro dos dados, baseados em suas similaridades. Este passo é vital para a compreensão da estrutura dos dados, permitindo que a Predição (aprendizagem supervisionada) utilize esses segmentos para construir modelos de Machine Learning mais focados e eficazes. Ao empregar os dados pré-processados e agrupados, o modelo de predição é capaz de aprender padrões mais detalhados e complexos, resultando em uma maior acurácia e poder de generalização nas estimativas futuras.

FIGURA 3.1 – Fluxograma geral do estudo

3.1 Dados e Modelagem (pré-processamento)

3.1.1 Base de Dados

Se utilizou a base no arquivo no formato *excel* "Banco GRU BSB 2018-21.xlsx", oriunda da pesquisa de satisfação do passageiro realizada nos aeroportos de GRU e BSB entre 2018 e 2021. No total, foram 39022 registros. Na aba Dados desse arquivo temos as seguinte variáveis que correspondem a identificação do passageiro:

Criou-se ainda uma coluna chamada *flag_pandemia* para indicar se o vôo foi realizado antes ou depois da pandemia do COVID-19, declarada em 11 de março de 2020 pela Organização Mundial de Saúde(OMS).

Além disso, tem-se os itens que correspondem a pesquisa de satisfação:

TABELA 3.1 – Variáveis de Identificação e Contexto da Pesquisa.

Variável	Descrição
AEROPORTO	Aeroporto onde se fez o embarque.
DIA_VOO	Dia em que se fez o embarque.
DIA_ENTREVISTA	Dia em que se fez a entrevista.
HORA_ENTREVISTA	Hora em que se fez a entrevista.
LOCAL_ENTREVISTA	Local em que se fez a entrevista.
NOME_PESQ	Nome da pesquisa à qual o registro pertence.
NOME_PAX	Nome do passageiro (Variável sensível).
CONTATO_PAX	Contato do passageiro (Variável sensível).
GENERO	Gênero declarado pelo passageiro.
IDADE	Idade declarada pelo passageiro.
ESCOLARIDADE	Escolaridade declarada pelo passageiro.
TIPO_VOO	Vôo doméstico ou internacional.
NUM_VOO	Número do vôo.
EMPRESA_AEREA	Empresa áerea.
QT_EMBARQUE	Quantidade de embarques nos últimos 12 meses.
DESTINO_FINAL	Local de Desembarque final do vôo.
TIPO_ACESSO	Meio de locomoção para chegar ao aeroporto.
CONEXAO_PAX	Indica se o vôo possui ou não conexão.
NOME_AUDIO	Código do áudio da pesquisa (para fins de auditoria).

TABELA 3.2 – Variáveis de Satisfação e Percepção dos Passageiros (Itens de Pesquisa).

Variável	Descrição do Item Avaliado	Escala
ITEM_8	Tempo de espera na fila de inspeção de segurança	1 a 5
ITEM_9	Organização do processo de inspeção de segurança	1 a 5
ITEM_10	Atendimento e cortesia dos funcionários da fila de inspeção de segurança	1 a 5
ITEM_11	Facilidade de encontrar caminho no terminal (sinalização)	1 a 5
ITEM_12	Disponibilidade de informações de voo	1 a 5
ITEM_13	Distância caminhada no terminal	1 a 5
ITEM_14_1	Qualidade de restaurantes e instalações para alimentação	1 a 5
ITEM_14_2	Variedade de restaurantes e instalações para alimentação	1 a 5
ITEM_15_1	Qualidade de lojas e estabelecimentos comerciais	1 a 5

Continua na próxima página

Tabela 3.2 – continuação

Variável	Descrição do Item Avaliado	Escala
ITEM_15_2	Variedade de lojas e estabelecimentos comerciais	1 a 5
ITEM_16	Disponibilidade de bancos, caixas eletrônicos e casas de câmbio	1 a 5
ITEM_17	Limpeza de banheiros	1 a 5
ITEM_18	Disponibilidade de banheiros	1 a 5
ITEM_19	Qualidade de rede sem fio e outras conexões de internet disponibilizadas pelo operador aeroportuário	1 a 5
ITEM_20	Disponibilidade de carrinhos para bagagem nas áreas públicas	1 a 5
ITEM_21	Conforto nas áreas de embarque	1 a 5
ITEM_22	Conforto térmico	1 a 5
ITEM_23	Conforto acústico	1 a 5
ITEM_24	Limpeza geral do aeroporto	1 a 5
ITEM_25	Facilidade para entrar ou sair de veículo na via de acesso junto à entrada do terminal (meio-fio)	1 a 5
ITEM_26	Disponibilidade de vagas de estacionamento	1 a 5
ITEM_27	Qualidade das instalações do estacionamento	1 a 5
ITEM_28	Relação preço/qualidade do estacionamento	1 a 5
ITEM_29	Relação preço/qualidade dos restaurantes	1 a 5
ITEM_30	Relação preço/qualidade das lojas	1 a 5
ITEM_31	Tempo de espera de check-in no aeroporto	1 a 5
ITEM_32	Eficiência do processo de check-in no aeroporto	1 a 5
ITEM_33	Atendimento e cortesia dos funcionários do check-in	1 a 5
ITEM_34	De forma geral, estou satisfeito com esse aeroporto	1 a 10
ITEM_35	O aeroporto atendeu minhas expectativas	1 a 10
ITEM_36	Este aeroporto se aproxima de um aeroporto ideal	1 a 10

3.1.2 Clusterização: preparação dos dados e escolha do número de clusters

Primeiramente,o fluxo foi executado em `scikit-learn`, com `random_state` fixo, e gráficos gerados em `matplotlib` configurando `set_xticks` com a faixa de k e `MaxNLocator(integer=true)`. Logo,Tendo o objetido de identificar perfis de passageiros a partir de variáveis *socioeconômicas e de viagem*, se excluiu os itens de satisfação.Assim, para a clusterização, se partiu *dataframe* original, e se fez as seguintes considerações:

1. **Colunas selecionadas**:

   `QT_EMBARQUE`, `ANO`, `IDADE`, `ESCOLARIDADE`, `TIPO_ACESSO` (numéricas); `AEROPORTO`, `CONEXAO_PAX`, `GENERO`, `flag_pandemia` (binárias); `EMPRESA_AEREA` (categórica).

2. **Binarizações e flags:**

   - `AEROPORTO`,`CONEXAO_PAX` e `GENERO` como 0/1 conforme dicionário;
   - `flag_pandemia` criada a partir de `ANO_MES`, com valor 1 para observações em período pandêmico ($\geq$ 2020-03) e 0 para o período anterior.

3. **Categórica de alta cardinalidade:** para `EMPRESA_AEREA` se extraiu as **Top-4** companhias por frequência no conjunto e mapeou-se as demais para a categoria "OUTRAS";

4. **Valores fantantes** se imputou por **mediana** (numéricas) e **moda** (binárias e categóricas). Quando relevante, a ausência pode ser acompanhada por *flags* binárias;

5. **Outliers:** Se agrupou todas as variáveis numéricas(com excessão da variável ANO, já tratada na criação da flag_pandemia) em faixas de valores para evitar outliers:

TABELA 3.3 – Codificação das Variáveis `IDADE` e `TIPO_ACESSO`.

Variável `IDADE`		**Variável** `TIPO_ACESSO`	
Código	**Faixa Etária**	**Código**	**Meio de Locomoção**
1	Até 18 anos	1	Aplicativos
2	18 – 25 anos	2	Carona
3	26 – 35 anos	3	Veículo Alugado/Ônibus
4	36 – 45 anos	4	Táxi
5	46 – 55 anos	5	Carro Próprio
6	56 – 65 anos	—	
7	66 – 75 anos	—	
8	76 anos ou mais	—	

TABELA 3.4 – Codificação das Variáveis `ESCOLARIDADE` e `QT_EMBARQUE`.

Variável `ESCOLARIDADE`		**Variável** `QT_EMBARQUE`		
Código	**Nível de Ensino**	**Intervalo**	**Código**	**Categoria**
0	Analfabeto	$[0, 1]$	1	Esporádicos
1	Ensino Fundamental	$[2, 6]$	2	Média Recorrência
4	Ensino Médio	$7+$	3	Alta Recorrência
6	Superior (Graduação)	—		
5	Espec. de Nível Superior	—		
2	Mestrado	—		
3	Doutorado	—		

6. **Escalonamento**: aplicou-se `MinMaxScaler` nas respectivas codificações das variáveis numéricas, levando-as ao intervalo $[0, 1]$ para compatibilizar escalas na distância euclidiana do K-Means.

A Tabela 3.5 resume o papel e as transformações aplicadas no pipeline de clusterização:

TABELA 3.5 – Variáveis utilizadas na clusterização (pré-análise) e respectivas transformações.

Variável	**Tipo**	**Papel**	**Transformação no pipeline**
QT_EMBARQUE	Numérica	Socioeconômico/Viagem	Imputação (mediana);Codificação;`MinMaxScaler`
IDADE	Numérica	Socioeconômico	Imputação (mediana);Codificação;`MinMaxScaler`
ESCOLARIDADE	Numérica	Socioeconômico	Imputação (mediana);Codificação;`MinMaxScaler`
TIPO_ACESSO	Numérica	Acesso ao Aeroporto	Imputação (mediana);Codificação;`MinMaxScaler`
ANO	Numérica	Temporal	Imputação (moda); `MinMaxScaler`
AEROPORTO	Binária	Controle (GRU/BSB)	Imputação (moda); conversão para `float`
CONEXAO_PAX	Binária	Viagem	Imputação (moda); conversão para `float`
GENERO	Binária	Socioeconômico	Imputação (moda); conversão para `float`
flag_pandemia	Binária	Temporal	Construída de ANO_MES; imputação (moda); `float`
EMPRESA_AEREA	Categórica	Viagem	Imputação (moda); mapeamento Top-4 + ”OUTRAS”; `OneHotEncoder`

Para o algoritmo de clusterização,aplicou-se para cada tipo, **K-Means** com inicialização **K-Means++** e distância euclidiana sobre a matriz transformada **X**. Os parâmetros principais(`init='k-means++'`, `random_state=42`; `n_init` e `max_iter`) foram preservados conforme padrão da biblioteca.

Na Escolha do número de clusters (k),se investigou $k \in \{2, \ldots, 10\}$ por dois critérios complementares:

- **Elbow Method**: para cada k, ajustou-se o K-Means e registrou-se a *inertia* (WCSS), com o intuito de observar o ponto de inflexão onde o ganho de WCSS se torna pequeno ("cotovelo").
- **Silhouette Score**: usando os resultados de K-Means de cada k, calculou-se o *silhouette* médio $\in [-1, 1]$ e plotou-se a curva Silhouette vs. k.

A decisão de k considerou a *faixa* sugerida pelo Elbow e, dentro dela, o k com maior (ou comparável) Silhouette médio, privilegiando soluções parcimoniosas e com boa coesão/separação. Soluções com clusters residuais muito pequenos foram evitadas por inspeção dos tamanhos relativos.

3.2 Predição da Satisfação com Random Forest

Primeiramente, todas as transformações e o estimadores foram encapsulados em `Pipeline` + `ColumnTransformer` +`random_state=42`, análogo à Clusterização. Assim, com o intuito de estimar a probabilidade de satisfação do passageiro e identificação de informações relevantes para a satisfação, reportou-se no conjunto de teste (AUC-ROC, Average Precision, F1, Precision, Recall), com base nos passos a seguir:

1. **Coluna Alvo:**

   A satisfação foi agregada em `SATISFACAO_MEDIA` como a média de ITEM_34,ITEM_35 e ITEM_36 e binarizada em `flag_satisfeito` conforme a escala 0–10: ≥ 8 para "satisfeito"(1) e "insatisfeito"(0) para o complementar. As linhas onde havia a falta de 1 resposta dos três itens se aplicou a média dos 2 presentes, nas linhas em que se havia a falta de 2 respostas se removeu da análise.

2. **Pesquisa Otimizada**:

   Com o intuito de remover correlação entre respostas de temas semelhantes, foi agrupado de acordo com a temática proposta no dicionário, indicado na tabela a seguir a construção detalhada de cada uma:

TABELA 3.6 – Construção das variáveis `V2_*`: média dos itens por temática.

Variável	Temática	Itens agregados
`V2_insp`	Inspeção de segurança	ITEM_8; ITEM_9; ITEM_10
`V2_cami`	Caminho	ITEM_11; ITEM_13
`V2_info`	Informações de voo	ITEM_12
`V2_rest`	Restaurantes	ITEM_14_1; ITEM_14_2; ITEM_29
`V2_loja`	Lojas/Retail	ITEM_15_1; ITEM_15_2; ITEM_30
`V2_banc`	Bancos/ATMs/Câmbio	ITEM_16
`V2_banh`	Banheiros/Limpeza	ITEM_17; ITEM_18
`V2_wifi`	Wi-Fi/Conectividade	ITEM_19
`V2_carr`	Carrinhos/Bagagem	ITEM_20
`V2_emba`	Áreas de embarque	ITEM_21
`V2_term`	Conforto térmico	ITEM_22
`V2_acus`	Conforto acústico	ITEM_23
`V2_limp`	Limpeza Geral	ITEM_23
`V2_mfio`	Meio-fio	ITEM_25
`V2_estac`	Estacionamento	ITEM_26; ITEM_27; ITEM_28
`V2_chkin`	Check-in	ITEM_31; ITEM_32; ITEM_33

As regras para cálculo das médias foi análoga a utilizada na criação da satisfação média.

3. **Variáveis escolhidas e Pré-Processamento:**

   - **Numéricas:** `V2_*`,`ANO`, `QT_EMBARQUE`, `IDADE`, `ESCOLARIDADE`, `TIPO_ACESSO`.
   - **Binárias:** `AEROPORTO`,`CONEXAO_PAX`, `GENERO`, `VOO_NACIONAL`, `flag_pandemia`.
   - **Categóricas:** `EMPRESA_AEREA`,`DESTINO_FINAL`,`ANO`

   O pré-processamento foi feito análogo a clusterização:implementou-se um `Pipeline` com `ColumnTransformer`, resumido ma tabela a seguir: Vale lembrar que todo ak de imputação/transformação ocorre apenas no treino para evitar *leakage*).

TABELA 3.7 – Preditores utilizados no Random Forest e respectivas transformações

Variável	**Tipo**	**Papel**	**Transformação no pipeline**
`AEROPORTO`;`CONEXAO_PAX`; `GENERO`; `VOO_NACIONAL`; `flag_pandemia`	Binária	Controles/Viagem	Imputação (moda); conversão para `float`
`QT_EMBARQUE`	Numérica	Demanda	Imputação (mediana); capping superior (P99)
`IDADE`, `ESCOLARIDADE`, `TIPO_ACESSO`	Numérica	Socioeconômico/Acesso	Imputação (mediana)
`V2_*_MEDIA`	Numérica	Experiência	Imputação (mediana)
`EMPRESA_AEREA`	Categórica	Viagem/Mercado	Imputação (moda); mapeamento Top–4 + "OUTRAS" ; `OneHotEncoder`
`DESTINO_FINAL`	Categórica	Rede/Rota	Imputação (moda);`OneHotEncoder`
`ANO`	Categórica	Temporal	Imputação (moda); `OneHotEncoder`

Dessa forma, utilizou-se **Random Forest Classifier** (CART binária), A base de dados é dividida para separar os conjuntos de treino e teste (80% e 20%), garantindo a reprodutibilidade. Aqui foram os argumentos utilizados:

- `test_size = 0.20`: 20% dos dados são reservados para teste.
- `stratify = y`: Mantém a proporção das classes da variável alvo (`y`) nos conjuntos de treino e teste (essencial para classificação).
- `random_state = 42`: Fixa a semente aleatória para garantir a reprodutibilidade dos resultados.

Definiu-se a grade de hiperparâmetros a serem exaustivamente testados para encontrar a combinação ótima do modelo *Random Forest* (`rf`).

3. Configuração do GridSearchCV e Validação Cruzada

Configuração do processo de otimização, utilizando busca em grade combinada com Validação Cruzada Estratificada para uma avaliação mais justa.

TABELA 3.8 – Configuração dos Parâmetros para GridSearch

Parâmetro	Valores Testados	Descrição
`rf_n_estimators`	`[200, 600]`	Define o número de árvores na Floresta Aleatória.
`rf_max_depth`	`[8, 20]`	Define a profundidade máxima que cada árvore pode atingir.
`rf_min_samples_split`	`[1, 2]`	Número mínimo de amostras necessárias para que um nó interno possa ser dividido.
`rf_min_samples_leaf`	`[1, 2]`	Número mínimo de amostras que um nó folha (terminal) deve ter.
`rf_max_features`	`["sqrt","log2"]`	O critério (número de *features*) a considerar ao procurar a melhor divisão.

TABELA 3.9 – Configurações de Controle do GridSearchCV

Parâmetro	Valor/Método	Descrição
`cv`	`StratifiedKFold (n_splits=5)`	Realiza a Validação Cruzada com 5 dobras (*folds*).
`scoring`	`"average_precision"`	A precição foi utilizada como métrica de avaliação usada para determinar o melhor conjunto de hiperparâmetros.
`n_jobs`	`-1`	O valor `-1` significa que todos os núcleos disponíveis serão usados para acelerar a busca.
`refit`	`True`	Após encontrar o melhor conjunto de hiperparâmetros, o modelo final deve ser retreinado automaticamente usando todos os dados de treino disponíveis.

Por fim, fez-se um estudo comparativo, com base nas métricas de avaliação do modelo, entre a predição com e sem a clusterização com base no melhor caso escolhido para cada tipo.

4 Resultados e Discussão

4.1 Resultados da Clusterização

Após o pré-processamento , a amostra válida para clusterização contém **38893** registros, com **27458: 71%** de GRU e **11435: 29%** de BSB. As variáveis utilizadas no processo foram as listadas na Tabela 3.5.

4.1.1 Elbow Method e Silhouette

Tem-se a seguir, a curvas WCSS para $k \in [2, 10]$. Além disso, a figura a seguir mostram o *Silhouette* médio por k:

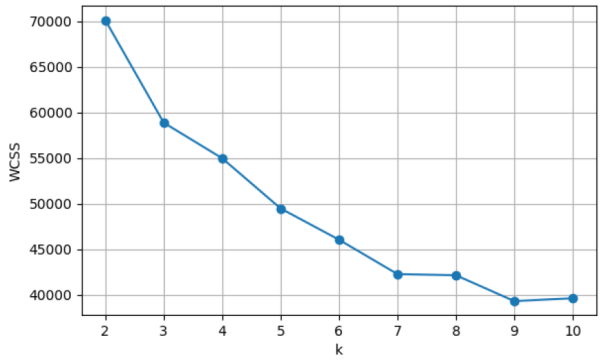


FIGURA 4.1 – Elbow Method para a base tratada

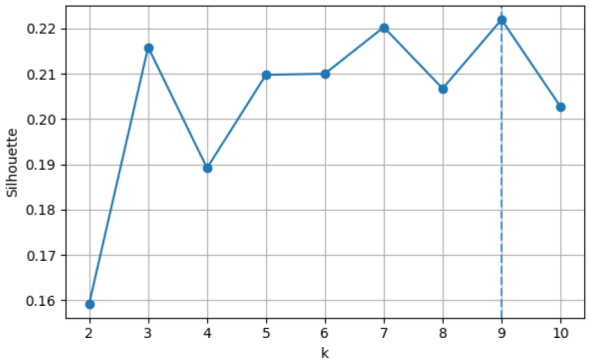


FIGURA 4.2 – Silhouette para a base tratada

4.1.2 Escolha de k

Combinando o *Elbow* e o *Silhouette*, adotou-se a regra: escolher um ponto de imflexão na faixa do cotovelo cujo *silhouette* seja máximo (próximo do máximo), tendo em vista a interpretação do resultado. Assim, se escolheu a seguinte lista de $k = 9$, tendo em vista a alta quantidade de variáveis escolhidas para realizar a clusterização.

4.1.3 Tabela dos Clusters

A seguir, temos uma tabela resumindo as principáis características de cada cluster, onde indica a representatividade das principais variáveis para cada cluster, bem como se essa característica apresenta alta variabilidade(Var.) ou predominância de algum valor(Pred).

TABELA 4.1 – Perfis dos Clusters

Variável	Cluster 0 (8.66%)	Cluster 1 (7.27%)	Cluster 2 (15.53%)	Cluster 3 (11.10%)	Cluster 4 (10.36%)	Cluster 5 (12.18%)	Cluster 6 (9.57%)	Cluster 7 (16.06%)	Cluster 8 (9.27%)
AEROPORTO	GRU (67.7%)	GRU (68.1%)	GRU (96.8%)	GRU (100%)	Pred.: GRU (55.6%)	GRU (100%)	GRU (61.6%)	GRU (61.0%)	BSB (100%)
GÊNERO	F (100%)	Pred.: M (55.7%)	Pred.: M (52.0%)	F (100%)	Pred.: M (56.1%)	M (100%)	M (100%)	Pred.: M (56.4%)	Pred.: M (59.8%)
IDADE	Var. (top: 3.0 28.9%)	Var. (top: 3.0 28.1%)	Var. (top: 3.0 29.0%)	Var. (top: 3.0 28.2%)	Var. (top: 3.0 28.8%)	Var. (top: 3.0 28.1%)	Var. (top: 3.0 27.1%)	Var. (top: 3.0 30.7%)	Var. (top: 4.0 27.4%)
ESCOLARIDADE	6.0 (67.8%)	6.0 (63.3%)	6.0 (71.6%)	6.0 (65.9%)	Pred.: 6.0 (58.9%)	6.0 (64.8%)	6.0 (68.7%)	Pred.: 6.0 (56.3%)	6.0 (73.1%)
EMPRESA_AEREA	GLO (100%)	ONE (100%)	Var. (top: AAL 9.6%)	TAM (82.8%)	GLO (100%)	TAM (81.0%)	GLO (100%)	TAM (92.2%)	TAM (79.8%)
QT_EMBARQUE	Pred.: 1.0 (46.4%)	Pred.: 1.0 (43.2%)	Pred.: 1.0 (58.2%)	Pred.: 1.0 (54.9%)	Pred.: 1.0 (50.6%)	Pred.: 1.0 (43.6%)	Var. (top: 2.0 38.1%)	Pred.: 1.0 (55.0%)	Pred.: 2.0 (41.7%)
DESTINO_FINAL	Var. (top: RJ 9.5%)	Var. (top: RJ 10.0%)	Var. (top: USA 23.7%)	Var. (top: BRA 8.9%)	Var. (top: PR 6.8%)	Var. (top: BRA 7.8%)	Var. (top: RJ 10.5%)	Var. (top: PR 6.6%)	Var. (top: SP 24.3%)
TIPO_ACESSO	Var. (top: CP 35.0%)	Var. (top: APP 38.9%)	Var. (top: APP 36.7%)	Pred.: CP (41.5%)	APP (97.5%)	Pred.: CP (43.6%)	Var. (top: CP 35.1%)	Pred: APP (98.9%)	Pred.: Carona (56.0%)
CONEXAO_PAX	N (99.8%)	N (72.4%)	N (92.7%)	N (99.8%)	S (100%)	N (99.9%)	N (99.8%)	S (100%)	N (99.8%)
VOO_NACIONAL	NAC. (91.6%)	NAC. (95.0%)	INT. (98.2%)	NAC. (74.9%)	NAC. (96.4%)	NAC. (77.5%)	NAC. (91.9%)	NAC. (91.8%)	NAC. (99.4%)
ANO	Pred.: 2019 (46.5%)	2018 (76.3%)	Var. (top: 2019 39.9%)	Pred.: 2019 (46.6%)	Pred.: 2019 (41.6%)	Pred.: 2019 (44.0%)	Pred.: 2019 (46.1%)	Var. (top: 2019 35.3%)	Var. (top: 2019 38.4%)
flag_pandemia	0 (81.2%)	0 (100%)	0 (80.8%)	0 (82.7%)	0 (72.6%)	0 (81.8%)	0 (81.6%)	0 (70.1%)	0 (73.3%)

Obs.: CP indica carro Próprio

Assim, tem-se a tabela a seguir resumindo os principais pontos distintos para identificação dos cluster

TABELA 4.2 – TABELA 4.2 – Resumo da Segmentação de Passageiros Aéreos por Cluster

Feature	Clusters de Destaque	Característica
Gênero (F)	0 (100%), 3 (100%)	Perfis exclusivamente Femininos.
Gênero (M)	5 e 6 (100%)	Perfil exclusivamente Masculino.
APP	4 (97.5%), 7 (98.9%)	Alta dependência de aplicativos para acesso.
Caror Próprio	0,3,5 e 6	Maior uso de Carro Próprio (CP).
Carona	8 (55.6%)	Perfil dependente de Carona/Outros.
Voos INT	2 (98.2%)	Viajante Internacional predominante.
Sem Conexão	0,2,4,5,6 e 8	Viagens simples, diretas ou com bilhete único sem conexão.
Com Conexão	4 e 7	Viagens com conexão.
Aeroporto Fixo	4 (GRU), 7 (BSB)	Base de partida fixa e definida.
Maior Embarque (12 meses)	5 (Pred. 2.0 - Média Recorrência)	Maior tendência a viagens recorrentes.

A distinção acentuada e complexa das características observadas nos clusters é uma consequência direta da natureza não supervisionada e aleatória da inicialização do algoritmo K-Means. Este método busca agrupar os dados com base na similaridade das features sem um conhecimento prévio das classes, resultando em grupos que podem misturar características de diversas colunas (idade, acesso, embarque, etc.). Uma abordagem alternativa para obter grupos com perfis mais puros e mais facilmente interpretáveis seria realizar uma segmentação prévia (estratificação) dos dados com base em uma variável-chave de alto impacto (por exemplo: Gênero, Voos Internacionais (INT/NAC) ou Empresa Aérea), e então aplicar o K-Means separadamente dentro de cada subgrupo.

4.2 Resultados da Predição com Random Forest

Após o processamento, análogo a clusterização, aplicou-se o algoritmo de Random Forest. O alvo `flag_satisfeito` foi definido por corte **TODO:** ≥ 8 na `SATISFACAO_-`

`MEDIA`. Os cenários comparados neste bloco *não usam* cluster como preditor; na próxima subseção será incluída a informação de cluster.

A Tabela **??** resume as métricas no teste para os diferentes clusters: **AUC-ROC**, **AUC-PR (Average Precision)**, **F1**, **Precision**, **Recall**, **Balanced Accuracy**, **MCC** e **LogLoss**. Observa-se que **TODO: destacar melhor e pior k e variações**.

TABELA 4.3 – Métricas de Avaliação de Desempenho e Satisfação por Cluster

Cluster	Accuracy	Precision	Recall	F1 - Score	ROC AUC	PR AUC	Satisfeitos (%)
0 (8.66%)	0.787	0.804	0.893	0.846	0.867	0.930	65.83
1 (7.27%)	0.769	0.775	0.893	0.830	0.848	0.906	63.00
2 (15.53%)	0.808	0.797	0.939	0.862	0.878	0.922	63.84
3 (11.10%)	0.793	0.786	0.935	0.854	0.863	0.913	64.75
4 (10.36%)	0.798	0.818	0.910	0.862	0.857	0.931	69.16
5 (12.18%)	0.785	0.789	0.856	0.821	0.870	0.903	57.64
6 (9.57%)	0.793	0.783	0.913	0.843	0.870	0.907	60.83
7 (16.06%)	0.823	0.834	0.933	0.881	0.883	0.943	70.17
8 (9.27%)	0.802	0.816	0.948	0.877	0.885	0.954	74.36
TOTAL (100%)	**0.806**	**0.809**	**0.921**	**0.861**	**0.879**	**0.928**	**65.55**

O desempenho dos modelo demonstrou um bom resultado geral, com um ROC AUC de 0.879 e um F1-Score de 0.861, indicando que o modelo tem uma boa capacidade de distinguir a satisfação. No entanto, o desempenho varia por grupo: os Clusters 7 e 8 apresentaram o melhor desempenho do modelo (maiores ROC e PR AUC), sugerindo que os perfis desses passageiros são os mais previsíveis em relação à sua satisfação.

Os Clusters 8 (74.36%) e Cluster 7 (70.17%) são os grupos mais satisfeitos. O Cluster 7 é o maior grupo, caracterizado por alto uso de APP de locomoção e voos com conexão, enquanto o Cluster 8 é o grupo do Aeroporto de Brasília (BSB) com predominância de acesso via carona e voos com recorrência média nos ultimos 12 meses, indicando que esses perfis de viagem e acesso tendem a ter maior satisfação.

Por outro lado, o Cluster 5 (57.64%) é o menos satisfeito. Este grupo, que se destaca por uma predominância do público masculino, embarque em GRU,predominantemente clientes da TAM e sem conexão.

Em suma, o Cluster 5 deve ser o foco primário para melhoria de serviços, enquanto os Clusters 7 e 8 representam os clientes mais felizes e previsíveis.

4.2.1 Variáveis mais influentes

Temos aqui, as variáveis mais importantes para verificar a predição dos passageiros para cada Cluster:

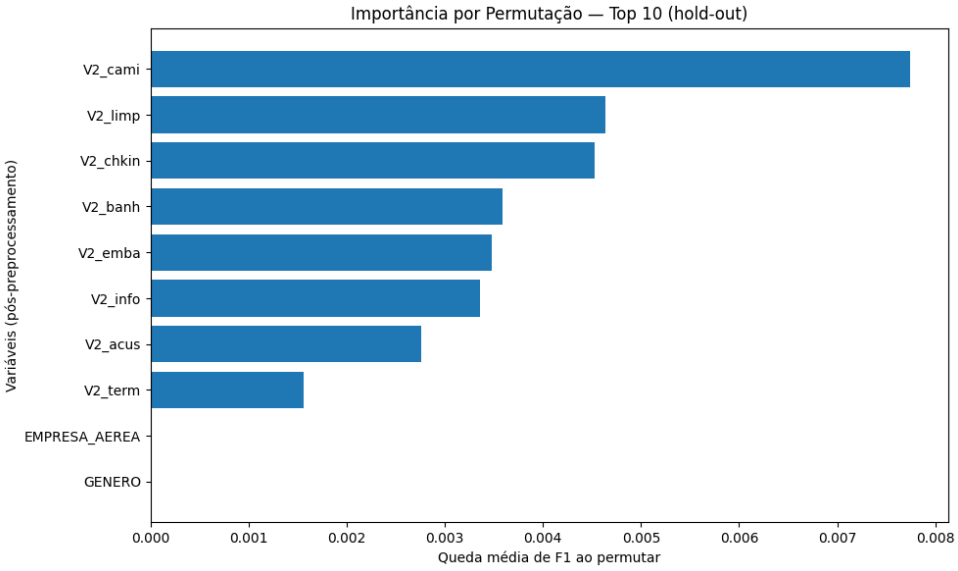


FIGURA 4.3 – Importância das variáveis na predição do Cluster 0

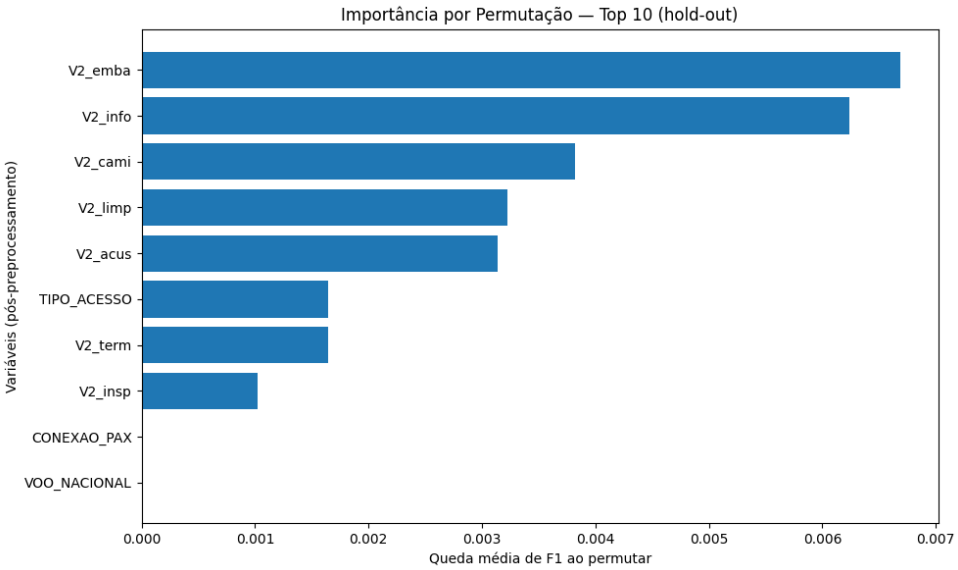


FIGURA 4.4 – Importância das variáveis na predição do Cluster 1

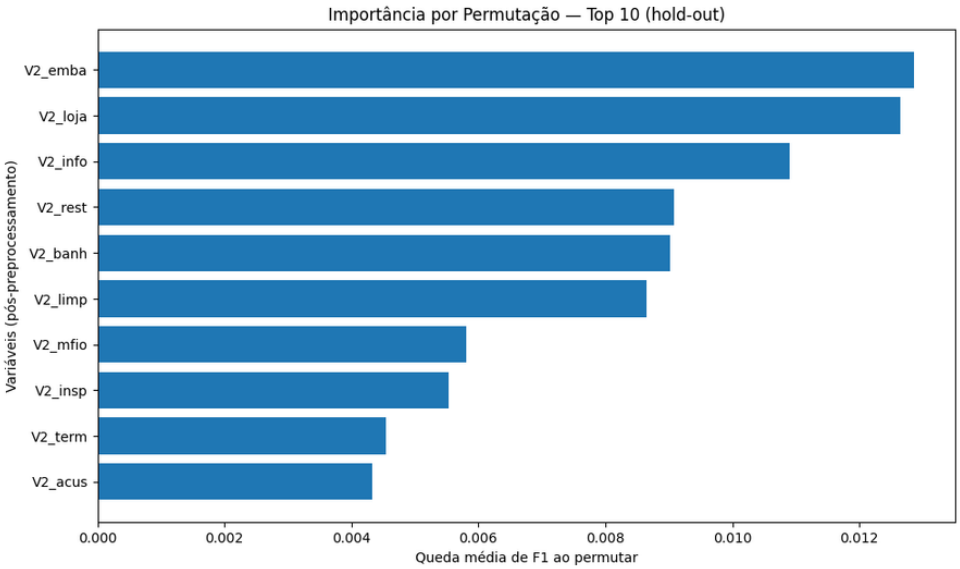


FIGURA 4.5 – Importância das variáveis na predição do Cluster 2

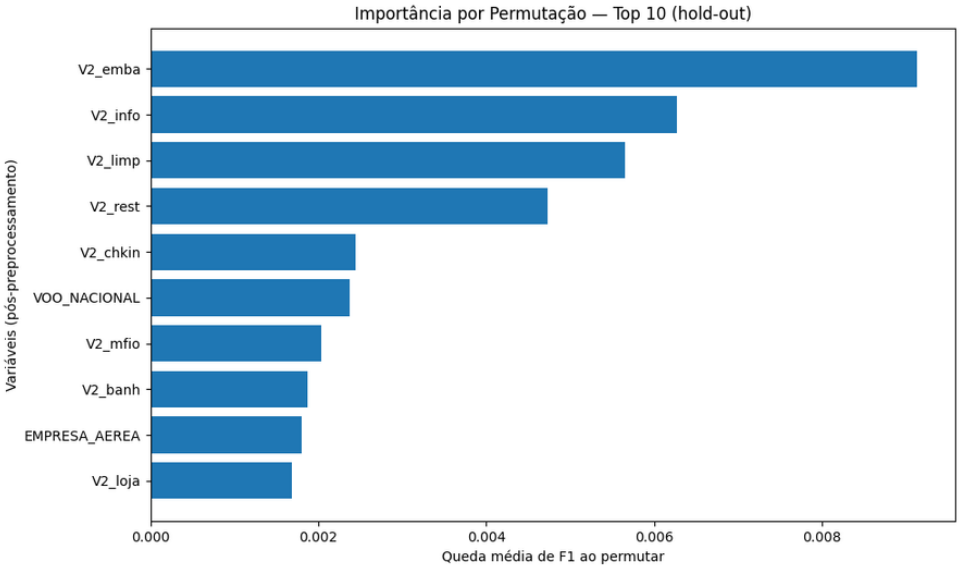


FIGURA 4.6 – Importância das variáveis na predição do Cluster 3

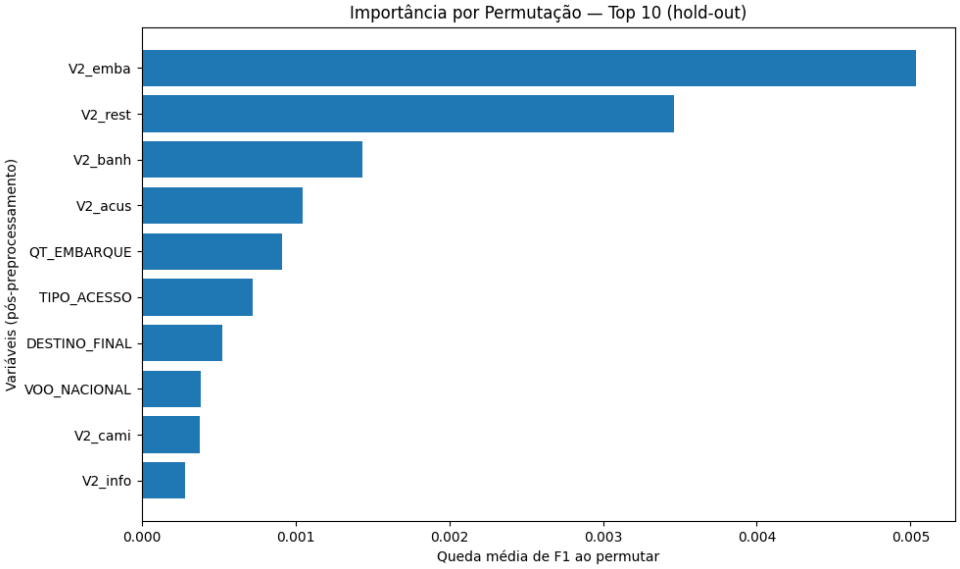


FIGURA 4.7 – Importância das variáveis na predição do Cluster 4

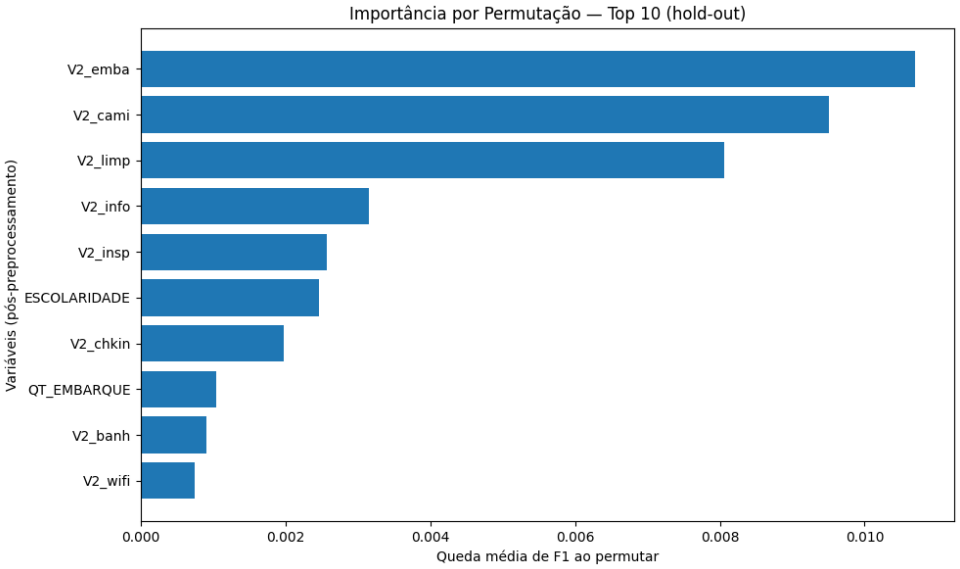


FIGURA 4.8 – Importância das variáveis na predição do Cluster 5

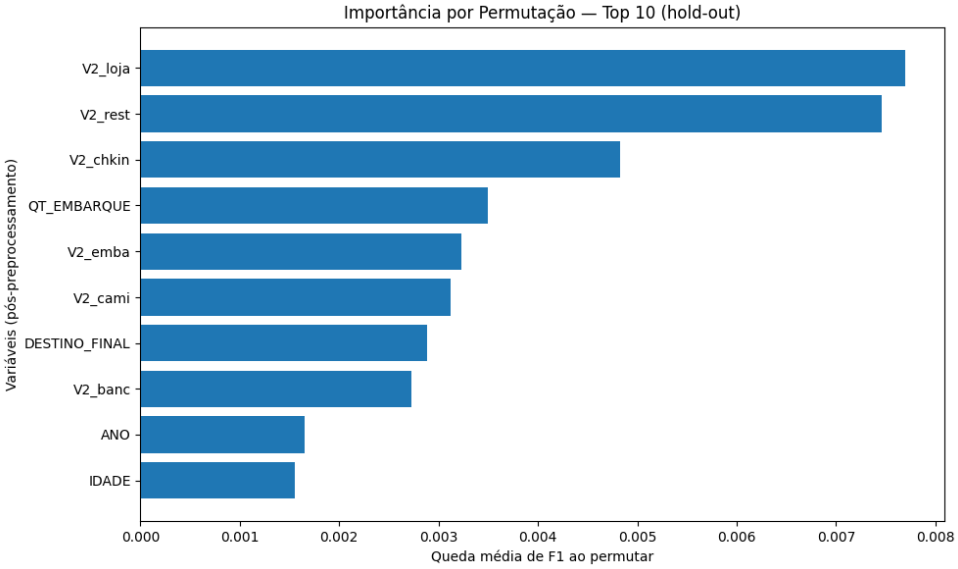


FIGURA 4.9 – Importância das variáveis na predição do Cluster 6

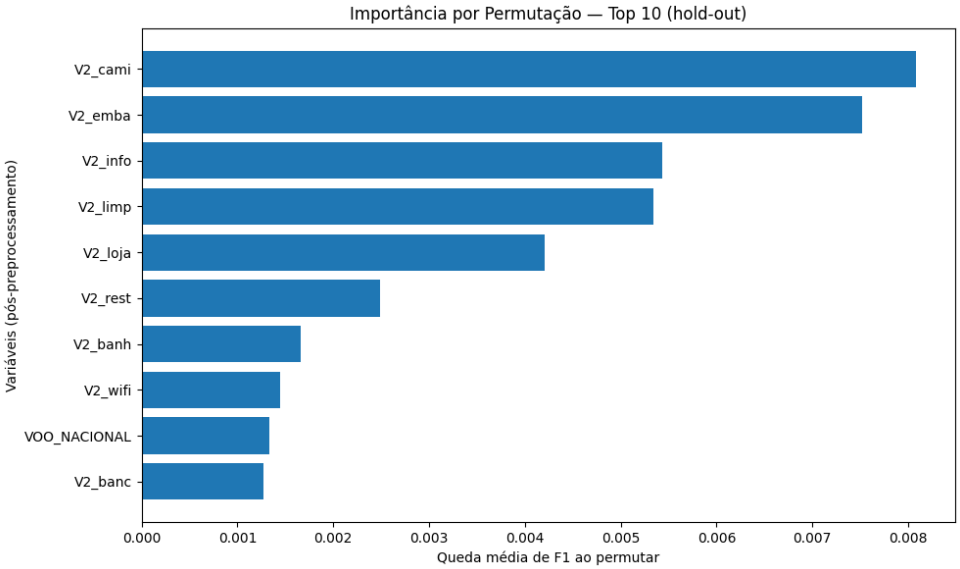


FIGURA 4.10 – Importância das variáveis na predição do Cluster 7

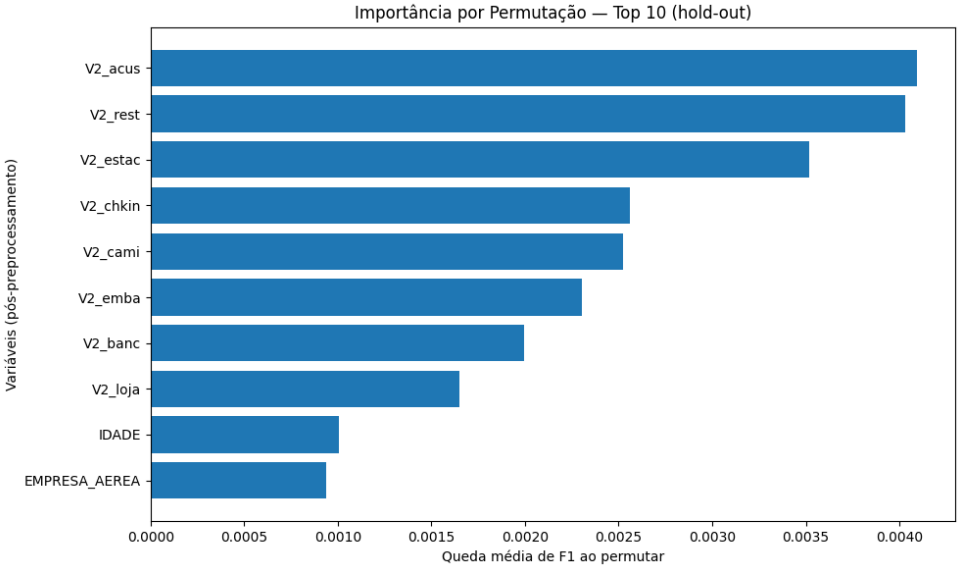


FIGURA 4.11 – Importância das variáveis na predição do Cluster 8

Assim, temos a seguinte tabela sintetizando o perfil do Cluster e sua satifação:

TABELA 4.4 – Análise de Clusters de Passageiros

Cluster	**Principais Características**	**Fatores Importantes**	**Satisfação**
0 (8.66%)	100% F, Carro Próprio, Sem conexão	Locomoção Interna ao Aeroporto	65.83
1 (7.27%)	Clientes ONE,Pré Pandemia	Embarque e Informações	63.00
2 (15.53%)	Sem conexão e Voos Nacionais	Embarque, Lojas e Informações	63.84
3 (11.10%)	Guarulhos, 100% F, Sem conexão	Embarque	64.75
4 (10.36%)	Gol, APP para locomoção e Com Conexão	Embarque e Restaurantes	69.16
5 (12.18%)	Guarulhos, 100% M e Sem Conexão	Embarque, Caminho e Limpeza	57.64
6 (9.57%)	100% M, Clientes Gol e sem conexão	Lojas e Restaurantes	60.83
7 (16.06%)	Clientes TAM, APP para locomoção, com conexão	Locomoção Interna e Embarque	70.17
8 (9.27%)	Guarulhos, Clientes TAM, Usam Carona, Sem conexão	Acústica, Restaurantes e Estacionamento	74.36

O grupo com satisfação mais elevada (Cluster 8: Passageiros TAM, Carona, Sem conexão), que valoriza fatores como Acústica, Restaurantes e Estacionamento, indicando que a infraestrutura e serviços de conforto são cruciais para a uma boa satisfação do passageiro. Em contraste, o Cluster 5 (100% M, Guarulhos, Sem Conexão) tem a pior satisfação, destacando a atenção para Embarque, Caminho e Limpeza como fatores determinantes para uma melhoria na procura da melhor satisfação desse público.

O fator Embarque é o recorrente como importante, afetando cinco clusters, o que o torna um requisito básico de qualidade. No entanto, para diferenciar e alcançar alta satisfação (acima de 70%), se observou que aspectos como Locomoção Interna (Cluster 7) e conforto ambiental/serviços (Cluster 8) se tornam diferenciais para obter uma maior satisfação média.

5 Conclusão

Primeiramente, a análise detalhada dos perfis permitiu identificar grupos majoritariamente separados por gênero, acesso ao aeroporto e a presençã ou não de conexão no vôo. Assim, aos gestores aeroportuários a capacidade de personalizar a experiência e focar melhorias com base nessas separações.

Em segundo lugar, a aplicação das técnicas de Machine Learning demonstrou ser altamente eficaz, com o modelo Random Forest alcançando 80.9% de precisão. Este desempenho confirma a boa capacidade dos modelos baseados em ML em capturar a complexidade não-linear da satisfação humana. A capacidade preditiva desenvolvida permite que o aeroporto aplique ações de melhoria a fim de atender às necessidades específicas de um grupo.

Como principal contribuição desse trabalho, tem-se o alinhamento da clusterização por K-Means para o agrupamento de passageiros para a posterior Predição. Uma vez que assim, se identifica grupos, que em primeira vista não são tão intuitivos e procura entender o que afeta a satisfação desse grupo.

Trabalhos futuros podem explorar a integração deste modelo preditivo com utilizando uma clusterização hierárquica, tornando o método do K-Means menos propenso a aleatoriedade, separando os grupos em questão previamente por gênero, região e voos internacionais.

Referências

ALMEER, M.; ALFAYEZ, A. **Measuring Airport Service Quality Using Machine Learning Algorithms**. [*S.l.*], 2022. Citado na p. 10.

BRASIL. SECRETARIA NACIONAL DE AVIAÇÃO CIVIL (SAC). **Pesquisa Nacional de Satisfação do Passageiro e Desempenho Aeroportuário**. [*S.l.*], 2024. Disponível em: https://www.gov.br/portos-e-aeroportos/pt-br/assuntos/noticias/2024/02/pesquisa-indica-alta-satisfacao-com-aeroportos-mas-aponta-para-necessidades-de-melhora-em-alguns-servicos/. Acesso em: 21 nov. 2025. Citado na p. 10.

CUNHA, R. W. S. da. **Machine Learning aplicado à Satisfação de Passageiros da Cia Aérea**. Curitiba, PR, 2023. Disponível em: `https://riut.utfpr.edu.br/jspui/handle/1/33666`. Citado na p. 15.

GOLDSCHMIDT, R.; BEZERRA, E.; PASSOS, E. **Data Mining: Conceitos, Técnicas, Algoritmos, Orientações e Aplicações**. 2. ed. Rio de Janeiro: Elsevier, 2015. Referência brasileira para a Matriz de Confusão, Acurácia, Precisão, Revocação e F1-Score. Citado na p. 15.

LLOYD, S. P. Least Squares Quantization in PCM. **IEEE Transactions on Information Theory**, v. 28, n. 2, p. 129–137, 1982. DOI: `10.1109/TIT.1982.1056489`. Citado na p. 12.

QIAN, J.; LI, Y.; WANG, Q. Aviation Passenger Segmentation through GANs Integrating K-Means Clustering: Innovating Airline Optimization. **International Journal of Advanced Information Science and Emerging Technology**, v. 6, n. 4, p. 141–155, 2024. DOI: `10.37391/IJAISER.060404`. Citado na p. 11.

ŞAHINBAŞ, K. Performance comparison of K-Means and DBSCAN methods for airline customer segmentation. **Black Sea Journal of Engineering and Science**, v. 5, n. 4, p. 158–165, 2022. DOI: `10.34248/bsengineering.1170943`. Citado nas pp. 10, 13.

FOLHA DE REGISTRO DO DOCUMENTO

1. CLASSIFICAÇÃO/TIPO	2. DATA	3. DOCUMENTO Nº	4. Nº DE PÁGINAS
TC	21 de Novembro de 2025	DCTA/ITA/TC-140/2025	36

5. TÍTULO E SUBTÍTULO:
ANÁLISE DE PERFIS SOCIOECONÔMICOS E PREDIÇÃO DA SATISFAÇÃO DE PASSAGEIROS EM AEROPORTO UTILIZANDO MACHINE LEARNING

6. AUTOR(ES):
Hermiro da Cruz Pessoa Junior

7. INSTITUIÇÃO(ÕES)/ÓRGÃO(S) INTERNO(S)/DIVISÃO(ÕES):
Instituto Tecnológico de Aeronáutica – ITA

8. PALAVRAS-CHAVE SUGERIDAS PELO AUTOR:
Clusterização, Satisfação, Passageiros, Aprendizado de Máquina, Random Forest

9. PALAVRAS-CHAVE RESULTANTES DE INDEXAÇÃO:
Aprendizagem (inteligência artificial); Indicadores socioeconomicos; Aeroportos; Qualidade de serviço; Passageiros; Algoritmos; Transportes.

10. APRESENTAÇÃO: (**X**) **Nacional** ( ) **Internacional**
ITA, São José dos Campos. Curso de Graduação em Engenharia Civil-Aeronáutica. Orientador: Prof. Dr. Alessandro Vinícius Marques de Oliveira. Publicado em 2025.

11. RESUMO:
Este trabalho apresenta uma análise aprofundada dos perfis socioeconômicos dos passageiros e o desenvolvimento de um modelo robusto de Machine Learning (ML) para predição de sua satisfação. A pesquisa foi conduzida utilizando uma pesquisa de satisfação de passageiros do Aeroporto Internacional de Guarulhos(GRU) e do Aeroporto Internacional de Brasília(BSB) entre os anos de 2018 e 2021, abrangendo variáveis como gênero, frequência de voo, escolaridade. A metodologia envolveu um pré-processamento de dados, uma clusterização com K-Means, seguida da implementação do algoritmo de Random Forest para ML. Os resultados demonstraram que o gênero, a forma de acesso ao aeroporto e a presença ou não de conexão são fatores importantes para o agrupamento de passageiros aéreos. O modelo de ML final, baseado em Random Forest, atingiu uma precisão de 80.6% (86.1% de F1-Score) na predição da satisfação. Além disso, concluiu-se que o embarque é um básico de qualidade. No entanto, para diferenciar e alcançar alta satisfação acima da média,se observou que aspectos como Locomoção Interna e conforto ambiental se tornam diferenciais.

12. GRAU DE SIGILO:
(**X**) **OSTENSIVO** ( ) **RESERVADO** ( ) **SECRETO**