Evasão Universitária na Universidade Federal Fluminense: Uma Análise de Agrupamento baseada em Aprendizado de Máquina

Mariana Moledo Moreira (UFF) - marianamoledomoreira@gmail.com
Gustavo Alexandre Sousa Santos (UFF) - gassantos@id.uff.br

Abstract. College dropout is a global challenge that impacts not only students but also higher education institutions and society. In this regard, this study focuses on understanding the profiles of students who drop out, aiming to develop prevention strategies and enhance the quality of higher education. By identifying clusters of students who have discontinued their courses using machine learning-based clustering techniques, the aim is to uncover the factors behind dropout. This analysis can improve academic programs and student support, promoting a more positive academic experience and retaining more students in college.

Keywords: High Education, Dropout, Educational Data Mining.

Resumo. A evasão universitária é um desafio global que afeta não apenas os estudantes, mas também as instituições de ensino superior e a sociedade como um todo. Nesse sentido, este estudo se concentra em compreender os perfis dos alunos que evadem, visando desenvolver estratégias de prevenção e melhorar a qualidade do ensino superior. Ao identificar agrupamentos de estudantes que abandonaram seus cursos por meio de técnicas de *Machine Learning* baseadas em clusterização, busca-se descobrir os fatores por trás da evasão. Essa análise pode aprimorar os programas acadêmicos e o suporte aos alunos, a fim de promover uma experiência universitária mais positiva e melhorar a permanência estudantil.

Palavras-chave: Ensino Superior, Evasão, Mineração de Dados Educacionais.

1. Introdução

A evasão universitária é um problema de grande relevância e impacto em todo o mundo, não apenas afetando o percurso acadêmico dos estudantes, mas também resultando em custos significativos tanto para as instituições de ensino superior quanto para a sociedade como um todo.

As Instituições de Ensino Superior (IES) enfrentam desafios significativos relacionados à evasão. De acordo com Silva Filho et al. (2007), as perdas de estudantes que iniciam, mas não terminam seus cursos são recursos públicos investidos sem o devido retorno no setor público, e no setor privado representam uma importante perda de receitas.

Em ambos os casos, a evasão é uma fonte de ociosidade de professores, funcionários, equipamentos e espaço físico.

1.1. Problema

A evasão no ensino superior representa um problema de grande magnitude, abrangendo dimensões sociais, acadêmicas e econômicas, e impondo sofrimento aos indivíduos que enfrentam a interrupção de seus projetos educacionais e profissionais (Borges, 2019). "A desistência em cursos universitários acaba por afetar diversos aspectos, uma vez que os alunos que abandonam a sala de aula frequentemente ingressam na sociedade despreparados para o mercado de trabalho" (CORDASSO et al., 2016, p.2).

Além disso, a evasão universitária exerce um impacto direto sobre os estudantes, interrompendo suas trajetórias educacionais e frequentemente prejudicando suas perspectivas de emprego. Essa situação pode ter efeitos adversos tanto em seu bem-estar econômico quanto emocional, criando obstáculos significativos para a realização de seus objetivos pessoais e profissionais.

Por último, mas não menos importante, a sociedade como um todo também sofre as consequências da evasão universitária. Egressos universitários têm o potencial de contribuir significativamente para o crescimento econômico, a inovação e o desenvolvimento social. Portanto, é crucial que as IES adotem uma abordagem estratégica, visando a conclusão dos cursos por parte dos alunos, capacitando-os a atingir suas metas e objetivos até o término de sua formação universitária (CORDASSO et al., 2016).

A evasão universitária é, sem dúvida, um dos problemas mais recorrentes e preocupantes do ensino superior, tanto no Brasil quanto em todo o mundo, afetando instituições públicas e privadas e causando efeitos em nível pessoal, na operação das instituições acadêmicas e no avanço da sociedade em sua totalidade. Quando estudantes abandonam prematuramente seus cursos, isso resulta em um déficit de capital humano, afetando a economia e a capacidade da sociedade de enfrentar desafios sociais e tecnológicos.

1.2. Objetivo

Nesse sentido, entender os perfis dos alunos que evadem é uma questão crucial no contexto educacional e esta pesquisa busca identificar e compreender os traços distintivos e as características dos estudantes que abandonam seus cursos universitários.

Analisar os perfis da evasão é fundamental para desenvolver estratégias eficazes na sua prevenção, permitindo que as IES atendam melhor às necessidades individuais de seus alunos e, assim, reduzam as taxas de abandono.

A compreensão sobre os estudantes que evadem pode fornecer *insights* valiosos para aprimorar os programas de orientação acadêmica e oferecer um suporte mais personalizado aos alunos em risco, contribuindo para uma trajetória universitária mais bem-sucedida e satisfatória. Portanto, investigar os perfis dos alunos que evadem é fundamental para melhorar a permanência estudantil e a qualidade do ensino superior.

Neste contexto, o presente trabalho tem como objetivo conduzir uma investigação abrangente, aplicando técnicas de Machine Learning (ML) baseadas em clusterização, a fim de identificar grupos (*clusters*) de alunos que evadiram de seus cursos na UFF. Ao

identificar esses grupos, espera-se descobrir quais fatores e padrões estão por trás da evasão. Isso pode ser super útil para as universidades melhorarem seus programas acadêmicos e o suporte aos alunos, o que, por sua vez, pode auxiliar em uma experiência acadêmica mais positiva.

Este estudo se baseia em pesquisas anteriores, incluindo o trabalho de Santos et al. (2020), que oferece *insights* valiosos sobre a evasão de estudantes universitários. Ao longo deste trabalho, será explorado os resultados obtidos por meio de técnicas de clusterização, destacando a importância da análise desses grupos identificados. O conjunto de dados abrange estudantes que ingressaram entre os anos de 2012 e 2014, bem como aqueles que desistiram ou se formaram até 2018 (SANTOS et al., 2020).

1.3. Proposta

No âmbito acadêmico e institucional, este estudo adquire um significado relevante, uma vez que a evasão universitária é um desafio urgente que afeta não apenas a UFF, mas também inúmeras outras instituições de ensino superior em todo o mundo. Compreender os motivos que levam os alunos a abandonarem seus cursos é fundamental para desenvolver estratégias eficazes de retenção, melhorar a qualidade dos programas acadêmicos e proporcionar um suporte mais adequado aos estudantes.

Este estudo ainda possui relevância social por tratar de um problema que impede o devido retorno dos investimentos públicos aplicados. Outro fato que marca a importância da pesquisa é o de poucas IES no Brasil possuírem programas bem elaborados e com resultados significativos que promovam a permanência nos cursos de ensino superior (Silva Filho et al., 2007).

Ao identificar os grupos de alunos que evadiram e os fatores subjacentes à evasão, este estudo oferece à UFF a oportunidade de adotar medidas proativas que não apenas beneficiam os alunos individualmente, mas também enriquecem a experiência acadêmica como um todo. Essa pesquisa contribui diretamente para a melhoria da qualidade do ensino superior e para a formação de profissionais mais qualificados, impactando positivamente a sociedade ao promover a retenção de alunos e o desenvolvimento de uma educação superior mais eficaz e inclusiva.

Além deste capítulo introdutório, este trabalho está estruturado em quatro seções principais: fundamentação teórica, método da pesquisa, resultados e análises, e por fim, as considerações finais.

2. Fundamentação Teórica

A definição da evasão varia de acordo com diferentes perspectivas. Santos (2014) a caracteriza como a descontinuação temporária do aluno que ingressou na Educação Superior após um ponto específico de seu percurso acadêmico. Já o Ministério da Educação (MEC, 1997) a compreende como a saída definitiva do estudante do curso de origem sem a conclusão, representando uma interrupção permanente. Vitelli e Fritsch (2016) a destacam como um processo complexo de exclusão influenciado por fatores internos e externos às instituições de ensino. Adicionalmente, Gaioso (2005) a descreve como uma interrupção no ciclo de estudos, enfatizando a necessidade de busca entender o fenômeno em sua totalidade, considerando todos os seus aspectos e contextos

relevantes. Portanto, a evasão universitária é um fenômeno social complexo que envolve tanto a interrupção temporária quanto a permanente dos estudos, sendo influenciado por uma variedade de fatores internos e externos às instituições de ensino.

A evasão universitária também é um fenômeno de alcance global e complexidade notável, despertando crescente preocupação entre acadêmicos e administradores educacionais em todo o mundo. Segundo Lorenzo-Quiles et al. (2023), esse problema se caracteriza por taxas significativas de desistência, afetando aproximadamente 20% dos estudantes que iniciam estudos de nível terciário, de acordo com dados da Organização para a Cooperação e Desenvolvimento Econômico (OCDE, 2019).

O impacto da evasão transcende fronteiras, sendo evidenciado em países como Malta, Espanha e Romênia, que registram altas taxas de evasão universitária, conforme revelado em um relatório do Eurostat (2020). Especificamente na Espanha, relatórios recentes do Ministério da Educação e Formação Profissional (MEFP, 2019) apontam que cerca de 30% dos estudantes abandonam as universidades espanholas, sendo essa evasão mais proeminente durante o primeiro ano de seus cursos. Esse panorama global da evasão na educação superior destaca a urgente necessidade de compreender e abordar esse desafio complexo que afeta não apenas os estudantes, mas também as instituições de ensino e a sociedade em geral.

No contexto brasileiro, a evasão universitária também é uma preocupação relevante e complexa que requer análise e intervenções cuidadosas. O Brasil enfrenta desafios específicos relacionados à evasão que estão intrinsecamente ligados à sua diversidade socioeconômica e geográfica. Uma das principais questões é a desigualdade socioeconômica que persiste no país, com muitos estudantes enfrentando dificuldades financeiras para custear seus estudos universitários. Essa desigualdade socioeconômica é agravada pela disparidade regional, onde algumas áreas do país têm um acesso mais limitado à educação superior de qualidade. Isso resulta em estudantes de diferentes regiões enfrentando desafios distintos ao buscar concluir seus cursos.

Para Moraes et al. (2006), a evasão no ensino superior abrange três tipos: evasão de curso, de instituição e de sistema. Suas causas internas incluem problemas na infraestrutura das universidades, atuação do corpo docente e falta de assistência socioeducacional. Causas externas envolvem decisões inadequadas em relação ao curso, dificuldades escolares e descontentamento com a futura profissão. Fatores socioeconômicos, como problemas financeiros e dificuldades em conciliar trabalho e estudos, são comuns na evasão, assim como problemas pessoais e distância entre domicílio e universidade.

Para Espíndola e Lacerda (2013), os cursos ministrados a distância oferecem aos alunos a vantagem da flexibilidade de horário e localização, ao mesmo tempo que apresentam desafios, incluindo a dificuldade em acompanhar as atividades propostas, a necessidade de autodisciplina e os obstáculos associados à tecnologia. Em sua pesquisa, os autores também observam que a maioria das evasões ocorre nos primeiros períodos do curso, sendo atribuída a fatores como a decisão de ingressar em outro curso superior, a falta de afinidade do aluno com o curso e problemas pessoais que surgem ao longo do curso.

Dados do Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP) indicam que a taxa de evasão em cursos de graduação no Brasil tem sido significativa. Em 2019, apenas um terço (33%) dos estudantes universitários concluíram

seus cursos dentro do período estimado. Esse índice aumenta para 50% quando se leva em conta os estudantes que graduaram em até três anos após o prazo esperado (INEP, 2019). Os motivos para a evasão no Brasil podem incluir dificuldades financeiras, falta de suporte acadêmico, falta de adaptação ao ambiente universitário e desafios socioeconômicos.

A fim de lidar com esse desafio, o governo do Brasil, instituições de ensino e organizações da sociedade civil têm adotado várias medidas, incluindo a implementação de programas de apoio aos estudantes, a criação de políticas de cotas para grupos historicamente menos representados e a ampliação da oferta de ensino superior em regiões remotas. Baggi et al. (2011) salienta que a avaliação institucional, devido ao seu envolvimento ativo na vida da instituição, possui uma perspectiva privilegiada da universidade. Ela desempenha um papel significativo ao contribuir para aprimorar os processos acadêmicos e administrativos, servindo como um instrumento essencial para ajustar metas e objetivos. No contexto da evasão escolar, a avaliação institucional pode desempenhar um papel proativo ao identificar procedimentos institucionais que possam prevenir a saída prematura dos alunos (Baggi et al., 2011).

Embora as taxas de evasão possam variar entre as instituições e os estados do país devido a essas disparidades, é amplamente reconhecido que a evasão é um problema generalizado que afeta muitos estudantes brasileiros. Isso não apenas limita o acesso à educação superior, mas também prejudica a capacidade do Brasil de desenvolver uma força de trabalho qualificada e promover o progresso social e econômico.

Dado esse contexto, compreender as causas e desenvolver estratégias eficazes para lidar com a evasão é fundamental para garantir que mais estudantes tenham a oportunidade de concluir seus estudos superiores e contribuir para o desenvolvimento social e econômico.

A capacidade de antecipar a evasão escolar nas instituições de ensino superior públicas desempenha um papel fundamental na elaboração de medidas que auxiliem no progresso educacional dos estudantes. Nesse contexto, as técnicas de ML surgem como aliadas valiosas para prever a evasão. A aplicação da tecnologia, abrangendo esses algoritmos e análise de *Big Data*, tem demonstrado ser uma abordagem altamente promissora na prevenção e previsão da evasão universitária.

As pesquisas conduzidas por Primão (2022), Santos et al. (2020), Moreira et al. (2020) e Lemos (2021) proporcionam insights enriquecedores sobre como essas técnicas vêm sendo empregadas para identificar padrões e fatores de risco associados à evasão. Ao analisar os dados educacionais, torna-se viável identificar indicativos precoces de desistência, tais como desempenho acadêmico deficiente, falta de engajamento ou desafios pessoais, viabilizando a intervenção proativa das instituições de ensino e a oferta de suporte personalizado aos alunos em situação de vulnerabilidade. Essa abordagem orientada por dados está gradativamente se consolidando como uma ferramenta imprescindível para que as universidades aprimorem suas estratégias de retenção de alunos, proporcionando, assim, uma experiência acadêmica mais satisfatória e efetiva.

3. Metodologia

Neste estudo, serão abordados os métodos aplicados na coleta e análise de dados, começando pela descrição da amostra, seguido do processo de coleta, análise descritiva e abordagem de ML adotada para realizar a clusterização.

3.1. Preparação dos Dados

O conjunto de dados aplicado nesta pesquisa contém informações detalhadas sobre os estudantes, abrangendo diversos aspectos, como as notas obtidas no exame de admissão universitária, o histórico acadêmico, os registros de políticas públicas, informações sobre raça-etnia e dados sociodemográficos.

É importante destacar que esses dados foram os mesmos utilizados na pesquisa intitulada "EvolveDTree: Analyzing Student Dropout in Universities" realizada por Santos, G. A. S.; Belloze, K. T.; Tarrataca, L.; Haddad, D. B.; Bordignon, A. L.; Brandao, D. N. Essa pesquisa foi apresentada na *International Conference on Systems, Signals and Image Processing* (IWSSIP) em 2020, realizada em Niterói, Brasil. Para uma análise mais aprofundada e informações específicas sobre o conjunto de dados, você pode consultar o artigo original no seguinte DOI: https://doi.org/10.1109/IWSSIP48289.2020.9145203.

Esse conjunto de dados abrange um total de 12.969 instâncias, representando estudantes de 106 cursos de graduação presenciais da Universidade Federal Fluminense. Uma descrição completa desses dados pode ser encontrada em https://github.com/gassantos/evolvedtree/blob/master/DATASETINFO.md. É válido ressaltar que esse conjunto de dados compreende estudantes que ingressaram entre os anos de 2012 e 2014, sendo estes evadidos ou concluintes até o ano de 2018.

O processo de preparação dos dados foi iniciado com uma filtragem por Status de Formação, na qual selecionou-se exclusivamente os alunos com o status "evadido". Em seguida, foi realizada a remoção de duplicatas, excluindo registros duplicados com base no identificador de aluno, assegurando que cada aluno seja representado apenas uma vez no *dataset*.

Inicialmente, especificou-se apenas os códigos de identificação dos cursos, carecendo das informações correspondentes aos nomes dos cursos. Para solucionar essa questão, foi adicionado ao conjunto de dados principal um arquivo que continha os detalhes de cada curso oferecido pela UFF. Esse processo permitiu associar as informações de nome do curso a cada aluno, fornecendo maior clareza sobre o curso ao qual cada estudante estava vinculado. Essa etapa foi essencial para enriquecer a análise.

Além disso, para ampliar a riqueza de informações, foi utilizada a categorização presente no arquivo disponível em http://www.coseac.uff.br/trm/2022/Arquivos/UFF-TRM2022-Anexo-14-Grupos.pdf. Isto possibilitou associar a área de estudo correspondente a cada um dos cursos, incrementando o nível informacional do conjunto de dados, o que é fundamental para uma análise mais abrangente e precisa.

3.2 Análise Exploratória

Através da análise exploratória dos dados, é possível identificar tendências significativas no perfil dos alunos que evadiram. Primeiramente, foi observado que a maioria dos alunos evadidos era solteira. Além disso, a idade de 23 anos destacou-se

como a mais frequente entre os alunos evadidos, sugerindo que esse grupo etário pode estar mais suscetível à evasão. No que diz respeito ao gênero, é notório que a maior parte dos evadidos era do sexo masculino, conforme apresentado na Tabela **1**.

Tabela 1 - Análise descritiva por Sexo dos alunos evadidos

Sexo	Frequência	Percentual (%)
M	5343	54,32%
F	4493	45,68%

Adicionalmente, a análise revelou que a maioria dos alunos que evadiram ingressaram no 1º semestre de cada ano letivo mediante a Tabela **2**. Essas descobertas também podem ser valiosas para o desenvolvimento de estratégias de retenção e apoio aos alunos, com o objetivo de reduzir as taxas de evasão no ensino superior.

Tabela 2 – Frequência por Semestre de Ingresso dos alunos evadidos

Semestre	Frequência	Percentual (%)
1	5574	56,67%
2	4262	43,33%

No primeiro momento, a variável "AREACURSO" foi submetida a uma análise de frequência, destacando a contagem de cursos em cada área. Esse procedimento permitiu uma visão detalhada da distribuição dos cursos em áreas específicas, revelando quais são as mais representadas na amostra. O gráfico de barras correspondente exibiu essa informação de forma visualmente eficaz, tornando evidente o número de cursos por área e facilitando comparações.

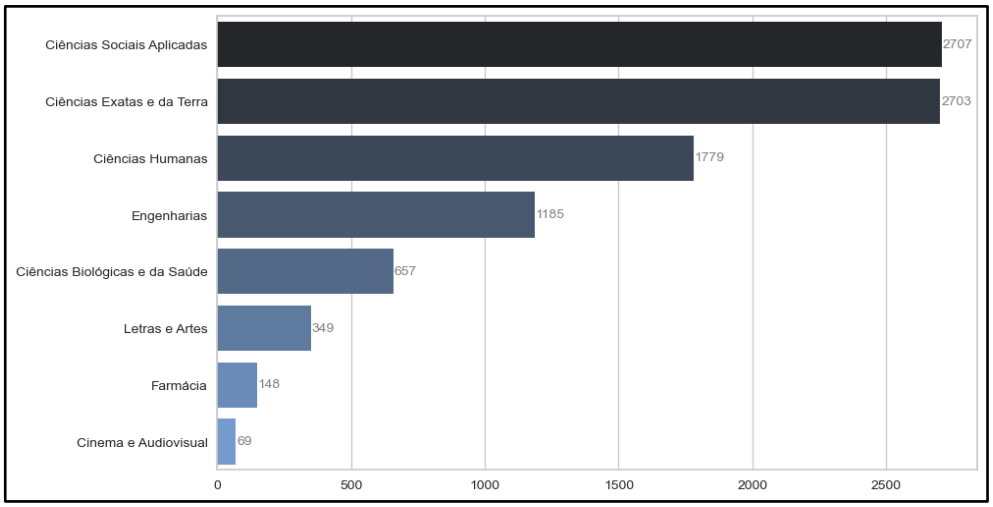

Figura 1 - Distribuição quantitativa de alunos por área de concentração

Além disso, uma etapa crucial dessa exploração envolveu a agregação das áreas de curso em categorias mais amplas, como "Exatas" e "Humanas". Essa categorização foi realizada com base nas áreas de conhecimento dos cursos, permitindo uma análise mais global das formações acadêmicas. O resultado foi a criação de uma nova coluna denominada 'Grupo_area_curso', que reúne os cursos em categorias mais abrangentes, conforme apresentada a Figura 2. O processo de contagem por grupo de área de curso exibiu o número de cursos em cada categoria e contribuiu para uma visão geral das áreas predominantes de estudo na instituição de ensino superior.

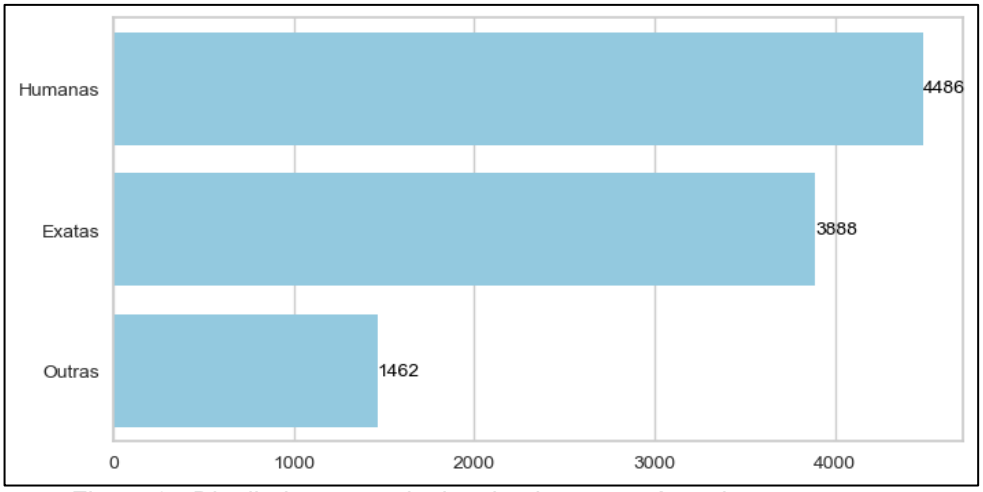

Figura 2 - Distribuição quantitativa de alunos por área de concentração

Também foi explorada a variável 'ACAOAFIRMATIVA', que descreve a presença ou ausência de ações afirmativas no processo de admissão. Da mesma forma, a Figura 3, apresenta uma distribuição quantitativa obtida pelo atributo 'Grupo_criterio'.

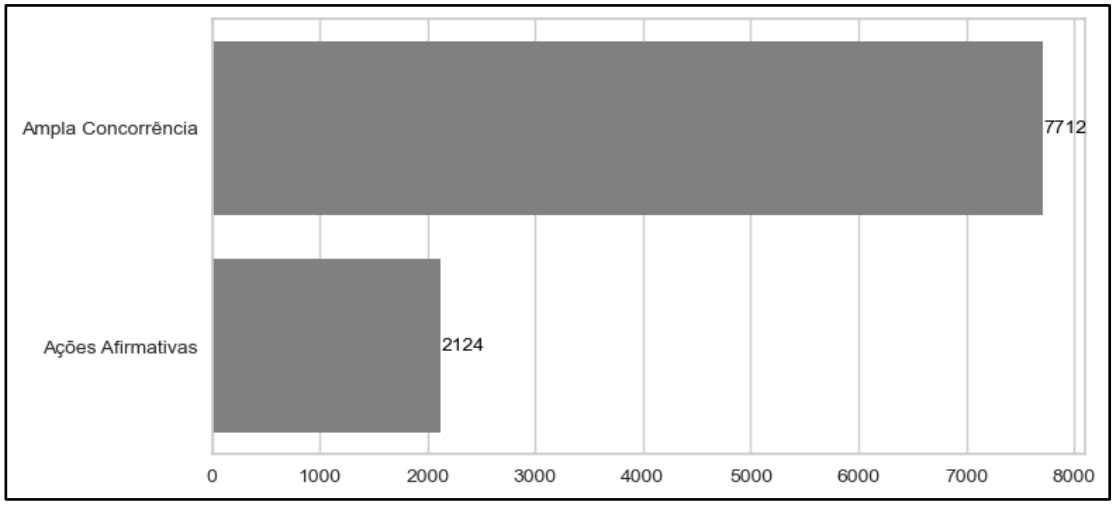

Figura 3 - Distribuição quantitativa de alunos por critérios de ações afirmativas

A contagem por grupo de critério foi calculada, a partir de uma função foi aplicada para agrupar as categorias em "Ampla Concorrência" ou "Ações Afirmativas", o que permitiu a análise de proporções de candidatos que se enquadram em cada uma dessas categorias distintas.

3.3. Pré-processamento dos Dados

No pré-processamento dos dados, adotaram-se duas técnicas importantes para melhor preparar o conjunto de dados. Primeiramente, a codificação *One-Hot* das variáveis categóricas. Isso foi aplicado às características, como 'Grupo_criterio', 'SEMESTREINGRESSO', 'COR', 'ESTADOCIVIL', 'SEXO' e 'Grupo_area_curso'. Essa codificação permite que variáveis categóricas sejam convertidas em um formato numérico, facilitando a aplicação de algoritmos de aprendizado de máquina. Em seguida, foi aplicada a normalização *Min-Max* às variáveis numéricas. Essa técnica dimensiona os valores numéricos para um intervalo entre 0 e 1, garantindo que todas as características tenham a mesma escala e evitando distorções nos algoritmos de aprendizado. Esses passos são fundamentais para garantir que o conjunto de dados esteja adequado à análise e modelagem subsequentes.

Essas etapas de pré-processamento desempenham um papel crucial na preparação dos dados para análise e modelagem. A codificação *One-Hot* permite que variáveis categóricas sejam representadas de maneira apropriada, enquanto a normalização Min-Max garante que as variáveis numéricas tenham uma escala consistente. Essas etapas foram fundamentais para garantir que todos os atributos estivessem em um formato compatível para a aplicação do algoritmo de clusterização.

3.4. Seleção de Variáveis

No âmbito da seleção de variáveis, implementou-se uma etapa crucial destinada a qualidade de dados. A estratégia de eliminação das características de baixa variância, foi desenvolvida por meio da classe *VarianceThreshold*, no qual, um limiar de variância de 0,21 foi estimado como critério para a obtenção das características mais relevantes. Essa etapa desempenhou um papel fundamental na redução da dimensionalidade dos dados, preservando apenas as variáveis relevantes. Ao descartar as características de baixa variância, foi possível atenuar o ruído presente nos dados e focar nos elementos de informação mais significativa, o que, por sua vez, resultou em um aprimoramento do desempenho algorítmico.

Após a conclusão da seleção, as variáveis que permaneceram no conjunto de dados incluem 'SEMESTREINGRESSO_1', 'SEMESTREINGRESSO_2', 'SEXO_M', 'SEXO_F', 'Grupo_area_curso_Humanas' e 'Grupo_area_curso_Exatas'. Essas variáveis representam o conjunto essencial de características que nortearão as análises posteriores.

3.5. Clusterização

O modelo *K-Means* é um algoritmo de aprendizado não supervisionado que tem como objetivo particionar um conjunto de dados em grupos (clusters) com base nas semelhanças entre os elementos. No contexto da nossa pesquisa, o *K-Means* foi usado em conjunto com a técnica do método do cotovelo, representada na Figura **4**, com o propósito de determinar o número ideal de clusters para agrupar os alunos.

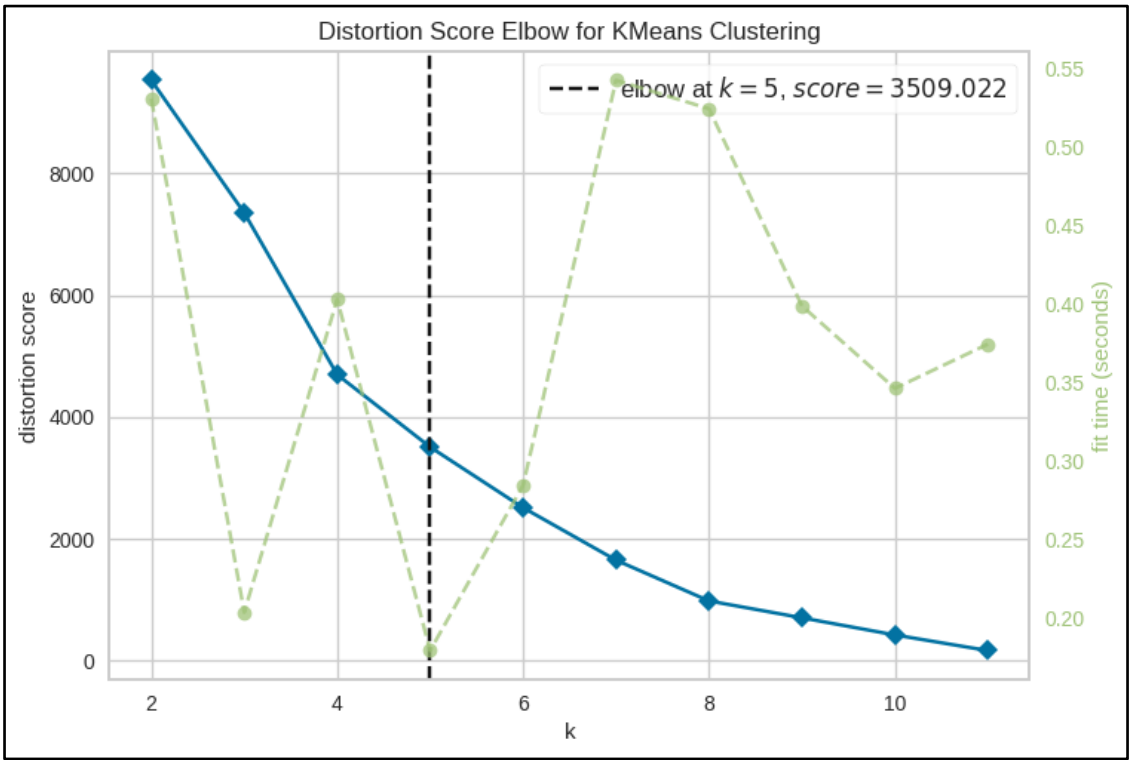

Figura 4 – Pontuação de distorção para o Agrupamento *K-means*

O método do cotovelo avalia a variabilidade explicada pelos clusters em relação ao número de clusters e nos ajuda a encontrar um ponto de equilíbrio onde acrescentar mais clusters não melhora significativamente a explicação das diferenças nos dados.

4. Análise dos Resultados

A compreensão aprofundada dos clusters é crucial para interpretar as tendências identificadas. A análise dos clusters é vital para compreender as particularidades de cada grupo e fundamentar decisões estratégicas. Para a ajudar na obtenção de explicabilidade, incorporou-se também a Árvore de Decisão, proporcionando uma representação visual e interpretável das regras orientadoras do modelo. A explicabilidade é crucial para garantir que os resultados sejam compreensíveis e válidos aos usuários finais, principalmente aos tomadores de decisão.

Na Figura 5 é possível verificar uma análise mais detalhada dos diferentes clusters identificados, numerados de 0 a 4, considerando variáveis específicas: "SEMESTREINGRESSO_1", "SEMESTREINGRESSO_2", "SEXO_M", "SEXO_F", "Grupo_area_curso_Humanas" e "Grupo_area_curso_Exatas". Cada cluster exibe um perfil distinto em relação a essas variáveis.

Cluster	Count	SEMESTREINGRESSO_1	SEMESTREINGRESSO_2	SEXO_M	SEXO_F	Grupo_area_curso_Humanas	Grupo_area_curso_Exatas
0	1909	0,00	1,00	0,00	1,00	0,52	0,30
1	2353	0,00	1,00	1,00	0,00	0,43	0,49
2	1238	1,00	0,00	1,00	0,00	1,00	0,00
3	2584	1,00	0,00	0,00	1,00	0,48	0,30
4	1752	1,00	0,00	1,00	0,00	0,00	0,80

Figura 5 - Análise descritiva dos clusters identificados

Cada cluster proporciona uma visão única das características dos seus integrantes, evidenciando variações significativas na composição com base nas variáveis analisadas. Essa análise detalhada é essencial para compreender as particularidades de cada grupo, fornecendo insights valiosos para embasar tomadas de decisões e estratégias direcionadas a essas distintas composições.

Uma outra forma de entender o perfil dos clusters é apresentada por meio deste mapa de calor na Figura **6**. O gráfico foi gerado com um tamanho de figura de 10 por 6 polegadas e utiliza a biblioteca *Seaborn* para visualização. Cada célula do mapa de calor representa uma estatística descritiva de um cluster específico em relação às variáveis consideradas. A coloração e os valores nas células refletem as tendências e diferenças em cada cluster.

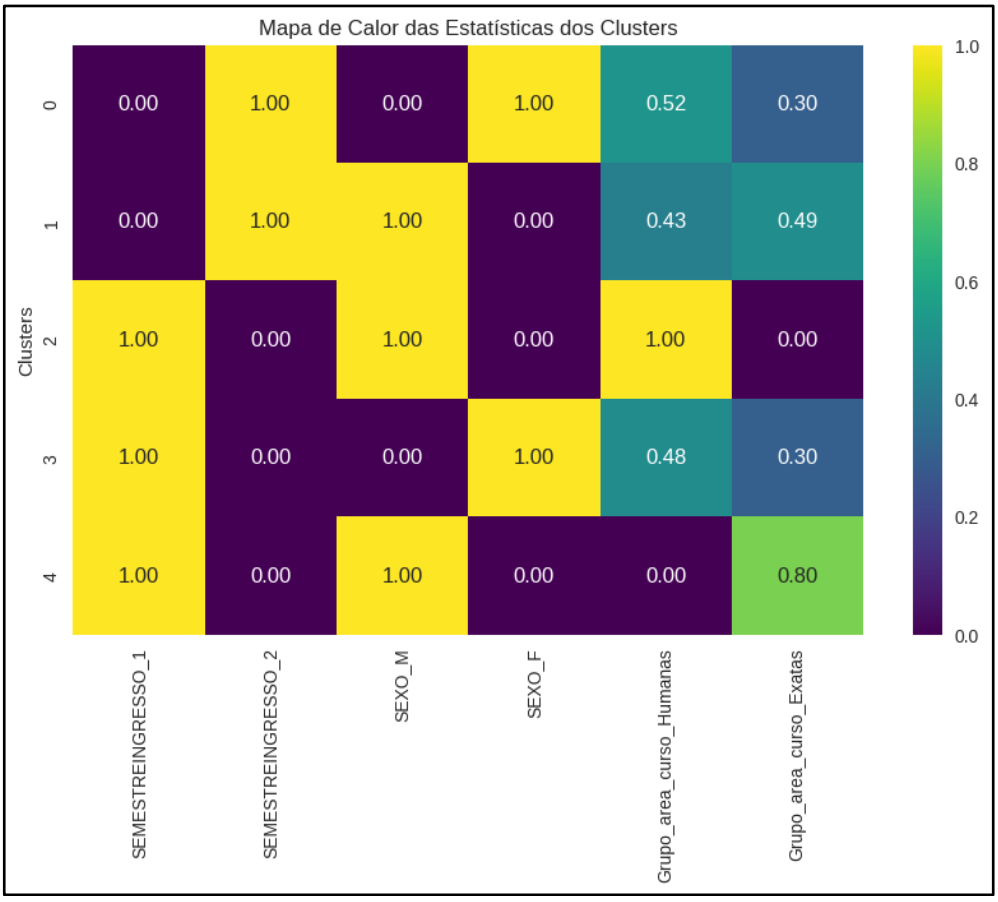

Figura 6 - Matriz de Correlação entre as variáveis e os clusters

Por exemplo, o Cluster 0, com média 0.0 em "SEMESTREINGRESSO_1" e 1.0 em "SEMESTREINGRESSO_2", indica que a maioria dos membros deste grupo ingressou no segundo semestre acadêmico. Além disso, este cluster é predominantemente composto por indivíduos do sexo feminino ("SEXO_F"). No que diz respeito à preferência de área de estudo, cerca de 52% dos membros pertencem à área de Humanas ("Grupo_area_curso_Humanas"), enquanto aproximadamente 30% escolheram a área de Exatas ("Grupo_area_curso_Exatas"). Essa representação gráfica é uma ferramenta valiosa para identificar visualmente padrões, relações e discrepâncias nos dados dos

clusters, contribuindo para uma compreensão mais profunda das características distintas de cada grupo.

Uma forma adicional de compreender os clusters é através da análise do *Silhouette Score*, que neste caso específico apresentou um valor de 0.61. Este resultado sugere que a segmentação dos dados em clusters foi bem-sucedida, pois o *Silhouette Score* se aproxima de 1. Isso indica que os pontos de dados estão bem próximos e coesos dentro de seus respectivos clusters, denotando uma clara separação entre os grupos. Um *Silhouette Score* tão próximo de 1 é um indicativo de uma alta qualidade na formação dos clusters, onde os pontos de dados dentro de cada cluster são mais similares entre si do que com aqueles pertencentes a outros clusters. Em resumo, o valor obtido é altamente positivo, indicando que o processo de clusterização foi satisfatório e que os dados foram agrupados de maneira distinta e bem balizada.

Neste estudo, utilizou-se também da Árvore de Decisão (AD) para entender como a classificação dos clusters foi realizada, já que a explicabilidade é uma abordagem valiosa na análise de dados. Essa abordagem complementar enriquece a compreensão dos padrões de segmentação dos dados, fornecendo *insights*. AD oferece uma representação visual e interpretável das regras que orientam as decisões do modelo, conforme apresentado na Figura 7. Quando aplicadas aos dados de cluster, essas árvores podem revelar quais características ou variáveis são mais influentes na separação dos dados em grupos distintos.

Cada nó na árvore representa uma decisão baseada em características específicas, permitindo auxiliar na identificação dos critérios que influenciaram a composição dos clusters. As ramificações indicam os caminhos que levam a diferentes grupos e isso pode identificar os critérios subjacentes que contribuíram na formação dos clusters.

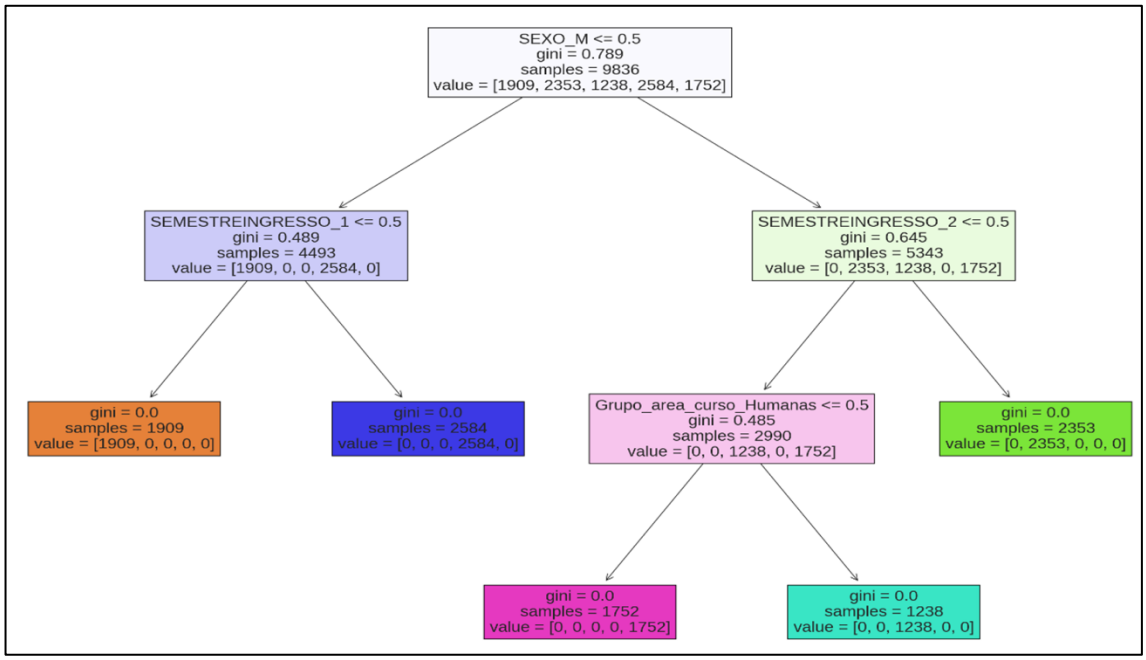

Figura 7 - Árvore de Decisão

Essa árvore de decisão ilustra como as características, tais como gênero, semestre de ingresso e grupo de área do curso, são utilizadas para classificar os dados em diferentes clusters. Cada nó da árvore representa um ponto de decisão com base em uma característica específica, e os ramos indicam os possíveis resultados. É uma ferramenta valiosa para entender o processo de classificação e como as características influenciam nas previsões.

5. Considerações Finais

A evasão é, certamente, um dos problemas que afligem as instituições de ensino em geral (Silva Filho et al., 2007). Nesta pesquisa, realizou-se uma extensa preparação dos dados, análise exploratória e pré-processamento, culminando na clusterização dos alunos com base em uma variedade de características. O conjunto de dados utilizado forneceu informações detalhadas sobre os estudantes, incluindo seu histórico acadêmico, informações demográficas e sociodemográficas, além de dados relativos a políticas públicas. Além disso, destaca-se que o conjunto de dados foi utilizado previamente em uma pesquisa anterior, o que reforça sua qualidade e relevância.

5.1. Contribuições do estudo

Durante a análise exploratória, foi possível identificar tendências nos perfis dos alunos que evadiram, incluindo a predominância de alunos do sexo masculino, solteiros e com idades em torno de 23 anos. A análise também revelou a importância do semestre de ingresso e permitiu a categorização das áreas de curso em "Exatas" e "Humanas".

O pré-processamento dos dados envolveu a codificação *One-Hot* das variáveis categóricas e a normalização *Min-Max* das variáveis numéricas, preparando o conjunto de dados para análises posteriores. A seleção de variáveis desempenhou um papel fundamental na otimização da qualidade do conjunto de dados, eliminando características de baixa variância e reduzindo a dimensionalidade.

A clusterização dos alunos com o algoritmo K-Means e a análise do *Silhouette Score* demonstraram a eficácia do processo de clusterização, resultando em grupos bem definidos e distintos. A representação visual dos clusters através de uma árvore de decisão destacou como as características, como gênero, semestre de ingresso e grupo de área do curso, influenciaram as previsões.

Essa pesquisa forneceu uma visão das características e dos perfis dos alunos evadidos, identificando fatores que podem estar relacionados à evasão. A análise dessas informações é essencial para o desenvolvimento de estratégias de retenção e apoio aos alunos, com o objetivo de reduzir as taxas de evasão no ensino superior. Além disso, as técnicas de pré-processamento e clusterização empregadas são valiosas para a análise de conjuntos de dados complexos e a compreensão das relações subjacentes.

5.2. Trabalhos Futuros

Embora este estudo se restrinja à investigação de uma única instituição de ensino superior e a uma ampla área de conhecimento, ele oferece uma valiosa contribuição para a pesquisa sobre evasão. Isso ocorre ao apresentar métodos e análises que podem ser aplicados em diferentes cursos e instituições de ensino superior. Essa pesquisa contribui

não apenas para a compreensão da evasão estudantil, mas também para a aplicação de técnicas avançadas de análise de dados em contextos educacionais.

Como continuação deste estudo, é promissor explorar abordagens adicionais que ampliem a compreensão da evasão de estudantes e melhorem as estratégias de retenção. Um caminho interessante seria a aplicação de outras técnicas de ML para avaliar outras técnicas de predição a fim de observar a evasão em outras perspectivas. Além disso, considerar a incorporação de novas variáveis e dados contextuais, como informações sobre o ambiente acadêmico, distância entre a residência do aluno e a Universidade e eventos específicos da instituição, pode enriquecer a análise e a capacidade de previsão. Essas abordagens poderiam contribuir para um entendimento mais abrangente e preciso da evasão de estudantes, permitindo a implementação de intervenções personalizadas e eficazes para a retenção acadêmica.

Referências Bibliográficas

BORGES, E. H. N. (2019). Modelos teóricos de análise da evasão no ensino superior aplicados à pesquisa sobre acompanhamento acadêmico dos discentes do setor público. Enfoques, 0(0), 83–95.

BRUNO, L. (1996). Educação, qualificação e desenvolvimento econômico. In: BRUNO, L. (Org.). Educação e trabalho no capitalismo contemporâneo: leituras selecionadas (pp. 91-123). São Paulo: Atlas.

CORDASSO, J. A. et al. Fatores Determinantes na Evasão de Acadêmicos no Ensino Superior: Estudo em um Município do Norte Mato-Grossense. 25 de novembro de 2016.

DIAS, E. C. M.; THEÓPHILO, C. R.; LOPES, M. A. S. Evasão no ensino superior: estudo dos fatores causadores da evasão no curso de Ciências Contábeis da Universidade Estadual de Montes Claros – UNIMONTES – MG. USP. Disponível em: https://congressousp.fipecafi.org/anais/artigos102010/an_resumo.asp?con=2&cod_trabalho=419&titulo=EVAS%C3O+NO+ENSINO+SUPERIOR%3A+ESTUDO+DOS+FATORES+CAUSADORES+DA+EVAS%C3O+NO+CURSO+DE+CI%CANCIAS+CONT%C1BEIS+DA

ESPÍNDOLA, R. M.; LACERDA, F. K. D. Evasão na Educação a Distância: um estudo de caso. EaD Em Foco, Rio de Janeiro, v. 3, n. 1, jun. 2013. Disponível em: https://eademfoco.cecierj.edu.br/index.php/Revista/article/view/174. Acesso em: 12 jul. 2020.

GAIOSO, N. P. L. O fenômeno da evasão escolar na educação superior no Brasil. 2005. 75 f. Dissertação (Mestrado em Educação) – Programa de Pós-Graduação em Educação da Universidade Católica de Brasília, Brasília, 2005.

INEP. (2019). Censo da Educação Superior 2018 - Notas estatísticas.

LEMOS, Í. V. R. Prevendo a evasão escolar em uma Instituição de Ensino Técnico utilizando Mineração de Dados Educacionais. 2021. 44 f. Trabalho de Conclusão de Curso (Graduação) - Universidade Federal Rural de Pernambuco, Bacharelado em Ciência da Computação, Recife, 2021.

LORENZO-QUILES, O.; GALDÓN-LÓPEZ, S.; LENDÍNEZ-TURÓN, A. (2023). Factors contributing to university dropout: a review. Frontiers in Education, 8, 1159864. DOI: 10.3389/feduc.2023.1159864.

MEC, Ministério da Educação. Comissão Especial de Estudos sobre a Evasão nas Universidades Públicas Brasileiras. Diplomação, retenção e evasão nos cursos de graduação em instituições de ensino superior públicas. Associação Nacional dos Dirigentes das Instituições Federais de Ensino Superior (ANDIFES), Associação Brasileira dos Reitores das Universidades Estaduais e Municipais (ABRUEM), Ministério de Educação e Cultura. Secretaria de Ensino Superior. Brasília, 1997. 152 p.

MORAES, J. O.; THEÓPHILO, C. R. Evasão no ensino superior: Estudo dos fatores causadores da evasão no curso de Ciências Contábeis da Universidade Estadual de Montes Claros - UNIMONTES. In: Universidade de São Paulo (Org.). Anais do Congresso USP de Iniciação Científica em Contabilidade, 2. Retirado em 13 junho 2007, de http://www.congressoeac.locaweb.com.br/artigos32006/370.pdf.

MOREIRA, F.J.R.; ALVES, M.A.Z. Aprendizagem de máquina na predição da evasão no ensino superior. 2020. Monografia (Especialização em Data Science e Big Data) - Universidade Federal do Paraná.

PRIMÃO, A. P. Uso de algoritmos de machine learning para prever a evasão escolar no ensino superior: um estudo no Instituto Federal de Santa Catarina. 2022. Dissertação (Mestrado Profissional) - Universidade Federal de Santa Catarina, Centro Sócio-Econômico, Programa de Pós-Graduação em Administração Universitária, Florianópolis, 2022. Disponível em: https://repositorio.ufsc.br/handle/123456789/238320. Acesso em: 2023.

SANTOS BAGGI, C. A. D.; LOPES, D. A. Evasão e avaliação institucional no ensino superior: uma discussão bibliográfica. Avaliação: Revista da Avaliação da Educação Superior (Campinas), v. 16, n. 2, p. 355-374, 2011. DOI: https://doi.org/10.1590/S1414-40772011000200007.

SANTOS, G. A. S.; BELLOZE, K. T.; TARRATACA, L.; HADDAD, D. B.; BORDIGNON, A. L.; BRANDAO, D. N., "EvolveDTree: Analyzing Student Dropout in Universities," 2020 International Conference on Systems, Signals and Image Processing (IWSSIP), Niterói, Brazil, 2020, pp. 173-178, DOI: https://doi.org/10.1109/IWSSIP48289.2020.9145203.

SANTOS, P. K. Abandono na Educação Superior: um estudo do tipo Estado do Conhecimento. Educação Por Escrito, Porto Alegre, v. 5, n. 2, p. 240-255, jul./dez. 2014.

SILVA FILHO, R. L. L.; MOTEJUNAS, P. R.; HIPÓLITO, O.; LOBO, M. B. C. M. (2007). A evasão no ensino superior brasileiro. Cadernos de Pesquisa, 37(132), 641-659. DOI: 10.1590/S0100-15742007000300007.