



UFOP – UNIVERSIDADE FEDERAL DE OURO PRETO
ICEB – INSTITUTO DE CIÊNCIAS EXATAS E BIOLÓGICAS
DECOM – DEPARTAMENTO DE COMPUTAÇÃO



LUCAS GOMES DOS SANTOS
VITOR HUGO LELES FONSECA

ANÁLISE DE DADOS EM OURO PRETO - MG

OURO PRETO

2023

LUCAS GOMES DOS SANTOS - 20.1.4108

VITOR HUGO LELES FONSECA - 17.2.4324

ANÁLISE DE DADOS EM OURO PRETO - MG

Relatório da Avaliação 02 da disciplina BCC325 - Inteligência Artificial, que consiste em análise de evasão por unidade acadêmica na UFOP - Universidade Federal de Ouro Preto no segundo semestre de 2021.

OURO PRETO

2023

SUMÁRIO

INTRODUÇÃO.....	4
OBJETIVO.....	4
METODOLOGIA.....	4
RESULTADOS.....	6
REGRESSÃO.....	8
VALIDAÇÃO CRUZADA.....	8
RANDOM FOREST.....	8
LIMITAÇÕES.....	9
CONCLUSÃO.....	9

INTRODUÇÃO

A análise de dados de evasão de alunos é uma questão crítica para as instituições de ensino superior, pode ter um impacto significativo não apenas na qualidade da educação oferecida, mas também nos recursos financeiros e na reputação da instituição. Para abordar esse desafio complexo, técnicas avançadas de análise de dados têm sido cada vez mais empregadas, incluindo abordagens de aprendizado não supervisionado, como o clustering.

Nesta abordagem, os dados coletados da Universidade Federal de Ouro Preto (UFOP), como informações de desempenho e evasão dos alunos, serão usados como base para o clustering. O objetivo é dividir os alunos em grupos com características semelhantes, permitindo que a universidade tome medidas proativas para intervir nos casos de maior risco de evasão. Isso pode incluir a implementação de programas de suporte acadêmico, intervenções psicossociais ou personalização do ensino para atender às necessidades específicas de cada grupo de alunos.

A aplicação do clustering na análise de dados de evasão, tem o potencial de fornecer insights valiosos para a gestão acadêmica. Isso pode resultar em estratégias mais eficazes para a redução da evasão, melhorando a experiência dos alunos e contribuindo para o sucesso geral da instituição.

OBJETIVO

O objetivo deste estudo é analisar os dados de taxas de evasão de alunos por unidade acadêmica, empregando uma abordagem de aprendizado não supervisionado através de clustering, seguida da construção de modelos de regressão para prever as taxas de evasão futuras. O final do trabalho implementamos o Random Forest para o caso de existir uma base de dados maior do que a implementada.

METODOLOGIA

Os dados utilizados referem-se ao segundo semestre de 2021 e foram obtidos através do Portal de Dados Abertos da UFOP, que tem como objetivo disponibilizar para toda a comunidade dados e informações públicas da Instituição, que podem ser usadas no desenvolvimento de aplicativos e ações. Foram utilizados 104 dados de alunos [evadidos](#) dos cursos de graduação presencial e à distância, e 1.823 dados de [disciplinas](#)

oferecidas pelos departamentos com informações sobre quantitativo de matriculados, aprovações, reprovações e trancamentos. Os passos incluem:

1. **Coleta e pré-processamento de dados:** Coletar e organizar os dados de evasão dos alunos da UFOP. Realizar a limpeza dos dados, tratando valores ausentes e outliers, e transformar as variáveis relevantes para a análise.
2. **Análise de Clustering:** Aplicar a técnica de clustering K-Means, para agrupar os alunos em diferentes segmentos com base em suas características. Isso ajudará a identificar padrões de evasão e a diferenciar grupos de alunos que apresentam comportamentos similares. O algoritmo K-Means será implementado utilizando bibliotecas de aprendizado de máquina, como Scikit-Learn.
3. **Análise Descritiva:** Realizar uma análise descritiva dos clusters formados, identificando as principais características de cada grupo. Isso permitirá compreender os perfis dos alunos mais suscetíveis à evasão e as diferenças entre os grupos.
4. **Modelo de Regressão:** Construir modelos de regressão linear para prever as taxas de evasão futuras com base nas variáveis relevantes identificadas na etapa anterior. Os modelos serão treinados separadamente para cada cluster, levando em consideração as particularidades de cada grupo.
5. **Avaliação e Validação:** Avaliar a qualidade dos modelos de regressão através de métricas apropriadas, como MSE (Mean Squared Error). Validar os modelos utilizando técnicas de validação cruzada para garantir que eles sejam capazes de fazer previsões precisas em dados não vistos.
6. **Análise de Resultados:** Interpretar os resultados obtidos, identificando quais variáveis têm maior influência na evasão dos alunos em cada cluster. Isso permitirá a formulação de estratégias específicas de intervenção para reduzir a evasão em diferentes grupos de alunos.

Unidades acadêmicas utilizadas:

1. Centro de Educação Aberta e a Distância
2. Escola de Direito, Turismo e Museologia

3. Escola de Farmácia
4. Escola de Medicina
5. Escola de Minas
6. Escola de Nutrição
7. Instituto de Ciências Exatas e Aplicadas
8. Instituto de Ciências Exatas e Biológicas
9. Instituto de Ciências Humanas e Sociais
10. Instituto de Ciências Sociais Aplicadas
11. Instituto de Filosofia, Artes e Cultura
12. Reitoria
13. Unidade Curso Sequencial

RESULTADOS

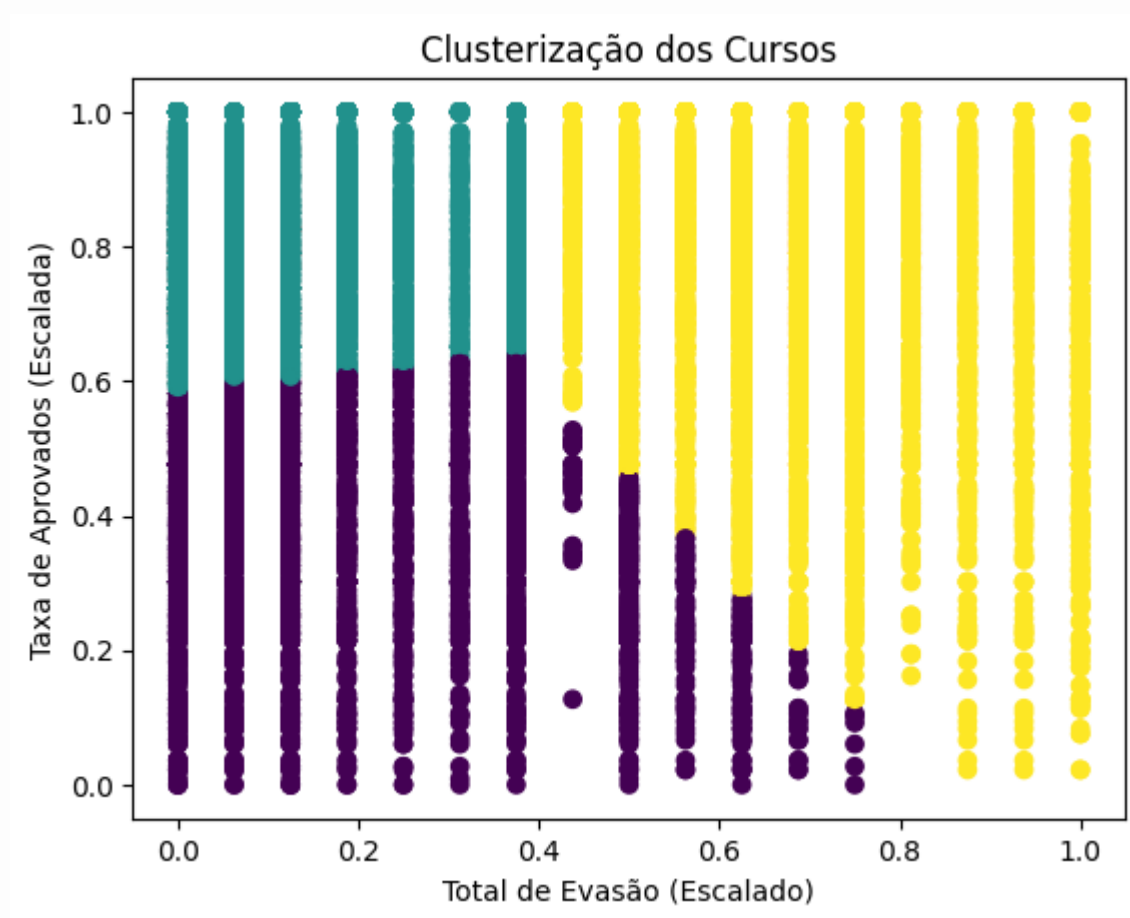


Gráfico 1: Resultado da clusterização

Roxo - cluster 0: Baixa a moderada evasão, baixo a moderado desempenho

Verde - cluster 1: Baixa Evasão, moderado a alto desempenho

Amarelo - cluster 2: Alta Evasão, moderado a alto desempenho

UNIDADE ACADÊMICA	TOTAL EVASÃO	DESEMPENHO
Escola de Farmácia	2,3893240983249800	43,663319167195100
Instituto de Ciências Exatas e Aplicadas	4,6470588235294100	37,543336128927900
Instituto de Ciências Exatas e Biológicas	3,5861895161290300	40,89058422783730

Tabela 1: Cluster 0

UNIDADE ACADÊMICA	TOTAL EVASÃO	DESEMPENHO
Centro de Educação Aberta e a Distância	2,7157894736842100	7,901586539622260
Escola de Nutrição	1,5393248731282400	8,458800276692120
Instituto de Ciências Humanas e Sociais	2,6019127316198400	78,725478225495
Instituto de Ciências Sociais Aplicadas	2,759656652360510	79,359163539596
Instituto de Filosofia, Artes e Cultura	2,8205574912891900	836744501242855

Tabela 2: Cluster 1

UNIDADE ACADÊMICA	TOTAL EVASÃO	DESEMPENHO
Escola de Direito, Turismo e Museologia	10,790633608815400	78,05205510584950
Escola de Medicina	8,8534032492112300	90,4340855668183
Escola de Minas	11,677469535908700	77,56709177240970

Tabela 3: Cluster 2

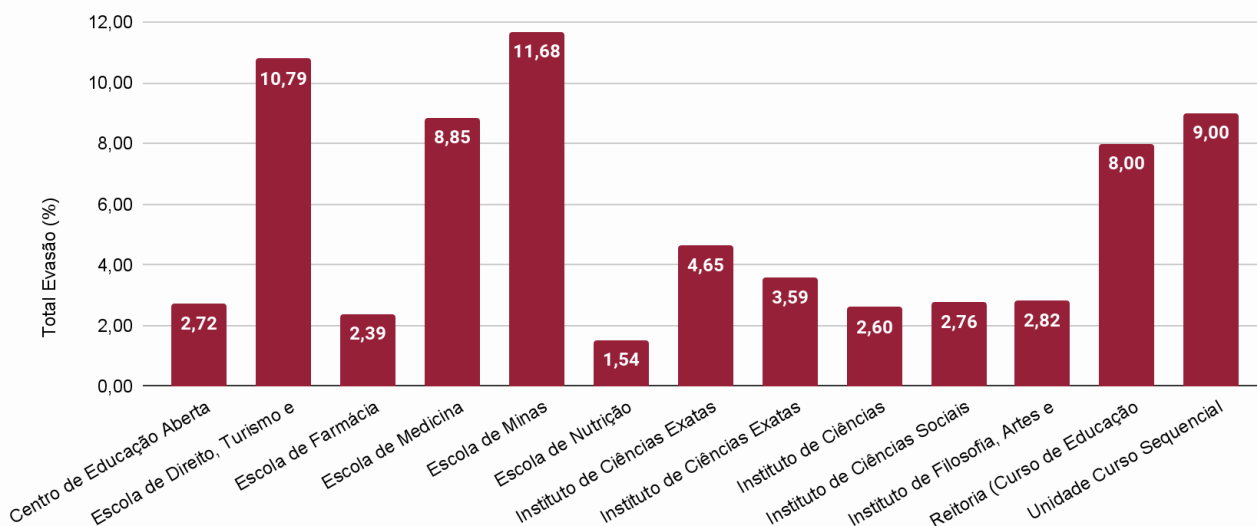


Gráfico 2: Porcentagem total de evasão por unidade acadêmica

REGRESSÃO

O Mean Squared Error (MSR), calculado como 20.9, representa o erro médio quadrado das previsões do modelo em relação aos valores reais da taxa de evasão. Interpretamos esse valor de erro como moderado e significa que o modelo não se ajustou muito bem aos dados, indicando que as previsões estão relativamente perto das reais.

VALIDAÇÃO CRUZADA

A validação cruzada foi realizada para avaliar o desempenho do modelo. O modelo apresentou um desvio padrão de cerca de 0.33, e um mean MSE de 21.43 e isso sugere que o modelo tem um desempenho consistente em diferentes conjuntos de dados de treinamento e teste, porém com um erro relativamente alto.

RANDOM FOREST

Usamos a Random Forest para calcular a importância das características (no caso, o desempenho) na previsão da taxa de evasão. O valor de importância 1 sugere que o desempenho é a única característica usada para fazer previsões.

Criamos essa implementação para fins educativos, de forma que, se obtivermos dados mais detalhados no futuro, não precisaremos recriar o código.

LIMITAÇÕES

É importante considerar que a taxa de evasão pode ser influenciada por uma variedade de fatores que não estão sendo considerados neste modelo, como características individuais dos alunos, características dos cursos, fatores socioeconômicos e outras variáveis não presentes nas características que usamos.

CONCLUSÃO

Essa implementação serve como um exemplo didático, fornecendo uma estrutura que pode ser expandida no futuro, caso tenhamos acesso a dados mais detalhados. A partir dessa experiência, entendemos a importância da seleção adequada de variáveis, do pré-processamento e da avaliação de desempenho dos modelos. Ao enfrentar problemas complexos como a evasão acadêmica, é essencial considerar múltiplos fatores e explorar abordagens mais avançadas para criar modelos mais eficazes.

Embora tenhamos obtido insights valiosos e tenhamos uma base sólida para análises futuras, reconhecemos que a complexidade da evasão acadêmica exige considerações mais abrangentes e a colaboração de diferentes áreas, como psicologia, sociologia e educação. Nosso trabalho nos proporcionou uma compreensão mais profunda da aplicação da análise de dados em contextos educacionais e abre portas para explorações mais detalhadas no futuro.

Apesar disso, fizemos a conclusão e análise em cima dos nossos dados.

O primeiro grupo representado por unidades acadêmicas que possuem baixa evasão e um baixo desempenho, como farmácia, exatas e biológicas, pode estar ligado a uma situação em que os alunos enfrentam dificuldades em suas áreas de estudo, possivelmente devido à complexidade das disciplinas. No entanto, a taxa de evasão baixa pode indicar que os alunos estão comprometidos em superar esses desafios, buscando apoio adicional e recursos de aprendizado.

O segundo grupo representado por unidades acadêmicas que possuem baixa evasão e moderado a alto desempenho, como nutrição, humanas, ciências sociais e filosofia, pode estar ligado a cursos que oferecem um ambiente mais propício para o engajamento dos alunos. O moderado a alto desempenho sugere que os alunos têm um bom nível de satisfação com o ensino e o ambiente acadêmico, o que pode estar

relacionado a abordagens de ensino inovadoras, interação com professores e colegas, e uma sensação de pertencimento.

O terceiro grupo representado por unidades acadêmicas que possuem alta evasão e moderado a alto desempenho, como Turismo, Direito, Medicina e Escola de Minas, pode estar ligado a cursos altamente exigentes, onde os alunos enfrentam uma carga acadêmica desafiadora. A alta taxa de evasão pode ser resultado da pressão acadêmica e emocional, juntamente com a necessidade de dedicar um tempo significativo ao curso. O moderado a alto desempenho pode indicar que os alunos que permanecem estão altamente motivados e dedicados a atingir seus objetivos.

Importante ressaltar que essas interpretações são baseadas em suposições gerais. Cada unidade acadêmica pode ter fatores específicos que contribuem para suas características de evasão e desempenho, e uma análise mais aprofundada seria necessária para obter respostas mais precisas.