# Trabalho Final Mineração de Dados

## Detalhes

Universidade Federal de Santa Maria (UFSM)
Centro de Tecnologia (CT)
Peso: 5
Disciplina de Mineração de dados (ELC1098)
Professor Dr. Joaquim Assunção
Autores: Bruno Perussatto & Gabriel Vinicius Schmitt Caetano

## Etapas do processo

### Seleção

Partindo da premissa que os dados foram previamente selecionados esta etapa foi ignorada.

### Pré-processamento

Para o pré-processamento, foram convertidos os arquivos para o formato CSV para padronizar e simplificar o processamento.
Em seguida foi identificado o arquivo “CE.csv” como duplicado e com nome fora do padrão, portanto foi removido do conjunto.
Então foi criado um arquivo único unindo todos os arquivos com a mesma estrutura, e no caso dos arquivos de Ingressantes e Formados também foi criada uma nova coluna “CENTRO” para garantir a consistência dos dados.

### Transformação

Para facilitar as análises a coluna “Situação” no conjunto de dados das turmas foi pivotado para uma coluna por situação disponível com a quantidade de alunos, assim como uma coluna por % de situação e a soma dos alunos e da %

A partir disso foi identificado uma inconsistência nos dados por alguns registros não alcançarem 100% dos alunos, indicando a falta ou falha no conjunto de dados
Sendo assim, os dados foram filtrados para considerar apenas os dados completos.

Outra transformação feita foi a conversão da coluna Ano + Semestre por uma coluna Ano com os valores fracionados para simplificar a plotagem, podendo usar em um eixo com os valores 2020, 2020.5, 2021, 2021.5, etc.

Também foi criada uma nova coluna Outros para agregar as situações diferentes de Aprovado/Reprovado, por conter quantidades menores e por ser uma informação menos relevante para a análise

Para facilitar a análise no conjunto de ingressantes e formados por centro, a coluna "ANO" foi convertida para o tipo numérico e o conjunto de dados foi filtrado para remover o valor "TOTAL". Em seguida, os dados foram agrupados por "CENTRO", "ANO" e "SEXO", e a soma dos ingressantes foi calculada. Foi realizada uma transformação de pivotamento para separar os ingressantes masculinos e femininos em colunas distintas.

Uma nova coluna de "TOTAL_INGRESSANTES" foi criada a partir da soma dos ingressantes masculinos e femininos, e o "PERCENTUAL_MASCULINO" foi calculado. A coluna "CENTRO" foi codificada numericamente para preparar os dados para o modelo de árvore de decisão.

## Análises gráficas e definição de hipóteses

Analisando os dados dos arquivos de Ingressantes e Formados, por meio de geração de gráficos de barras utilizando o python. O caminho escolhido foi gerar informações sobre a distribuição de sexo por centro, distribuição de ingressantes e formados por ano em cada centro e a distribuição geral de ingressantes e formados por todos os centros.
Na Figura 1 pode ser visto imagens de alguns gráficos utilizados para o embasamento da hipótese analisada em nosso trabalho (as demais imagens estão no pdf de apresentação ou geradas via script que será enviado):

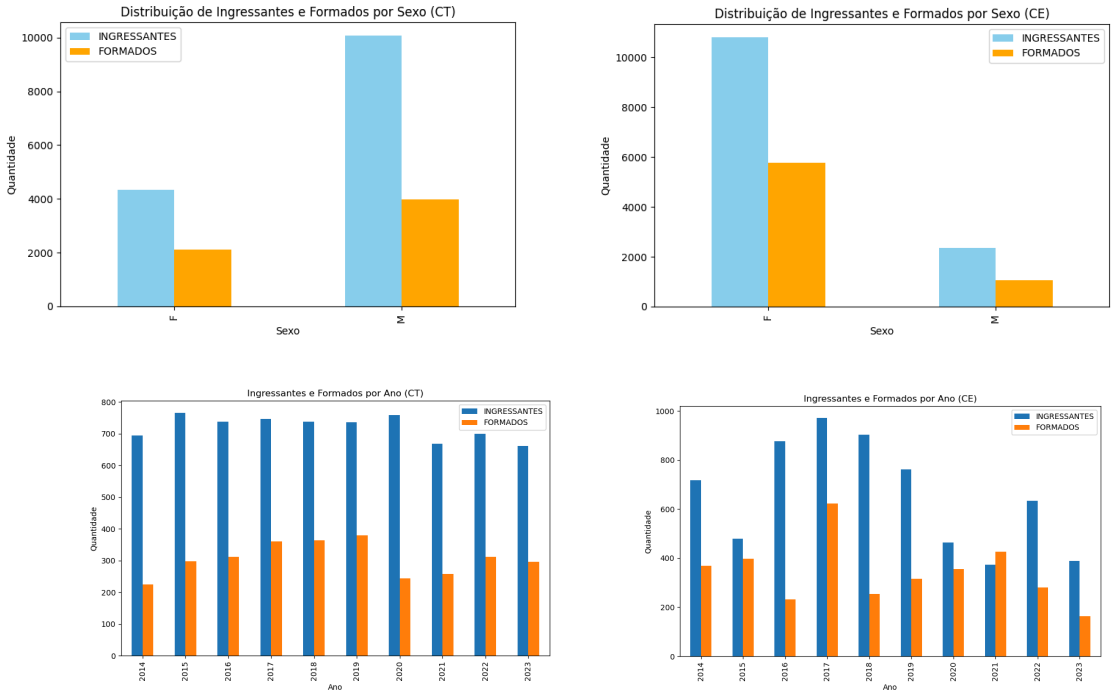


Figura 1

Além disso também foram gerados gráficos para visualizar a taxa de aprovação, reprovação e outras situações em cada disciplina, para compreensão melhor dos dados e das informações disponíveis, conforme pode ser visto na figura 2

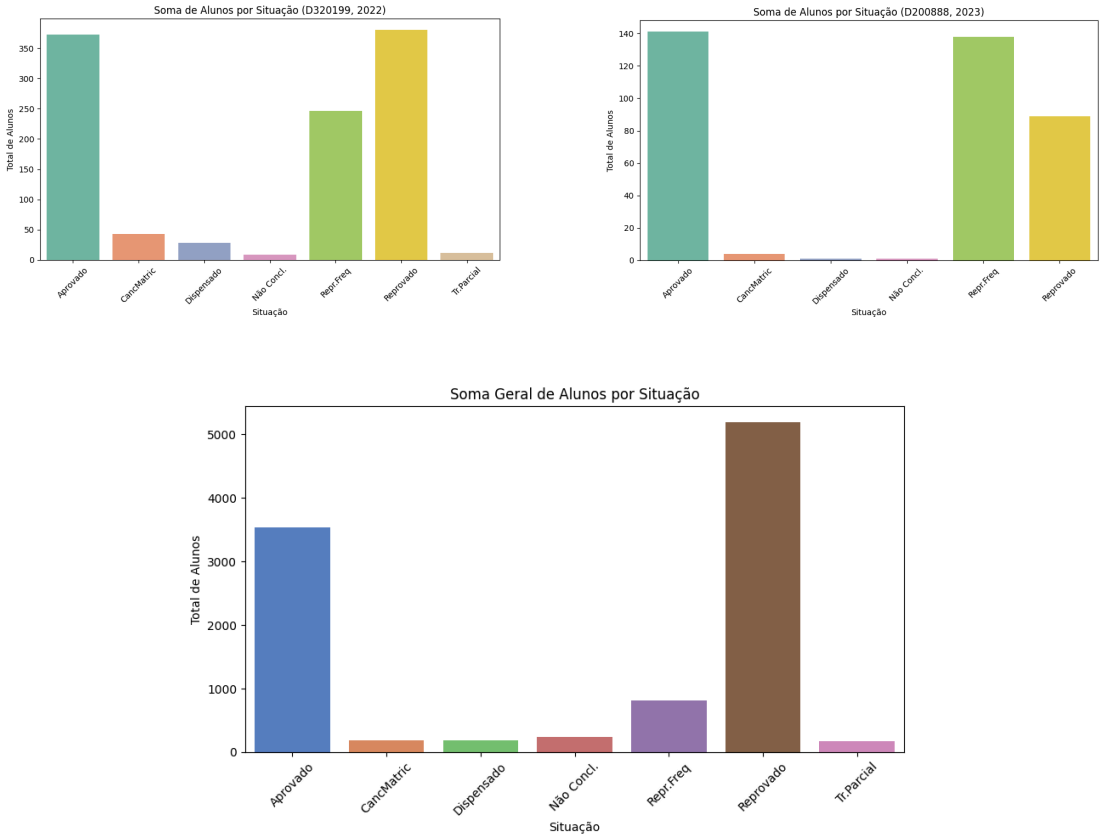


Figura 2

Outra visualização criada foi dos desempenhos das turmas para cada professor no decorrer do tempo. conforme a Figura 3

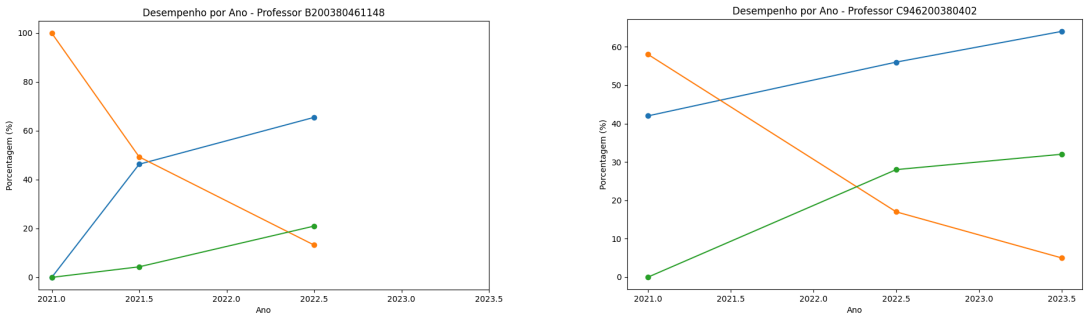


Figura 3

Nestes gráficos é possível identificar não apenas o desempenho individual das turmas, mas também o progresso do professor, o que pode ser usado para encontrar fatores que auxiliaram os professores a obterem resultados melhores e assim promover uma qualidade maior do ensino na instituição

## Mineração de dados

Com base nos dados tratados conforme descrito anteriormente, foi possível realizar as seguintes análises a fim de buscar informações relevantes.

## - Apriori

Buscamos identificar algum padrão de associação que pudéssemos prever informações de aprovação ou reprovação, relacionadas com dados das disciplinas ou dos semestres. Assim foram gerados os Heatmaps das regras encontradas conforme a Figura 3

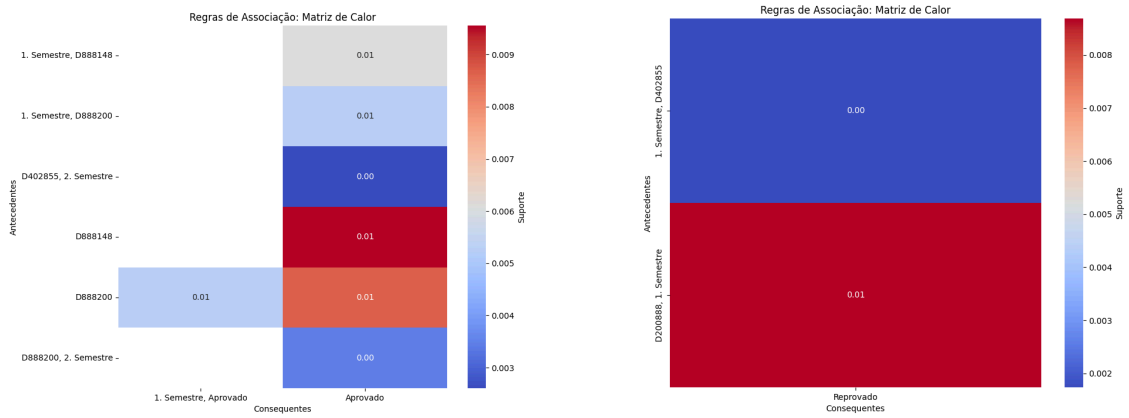


Figura 4

Esta análise encontrou apenas algumas relações fracas entre as disciplinas D200888 e D402855 no primeiro semestre com resultados de reprovação
E algumas relações um pouco mais fortes entre as disciplinas D888148, D888200 com resultados de aprovação.

## - Classificação (Árvore de Decisão):

O objetivo do modelo é identificar padrões que permitam prever a distribuição de alunos por centro com base em características como a porcentagem de ingressantes de um determinado sexo, o número total de ingressantes e o ano de entrada.

Para isso, foram analisados gráficos históricos que indicam que o número de ingressantes por ano supera o de formados e que há uma tendência de predominância de um sexo em certos centros. A partir desses padrões, foi criada uma árvore de decisão para classificar os alunos conforme essas variáveis, com foco na previsão do centro de origem.

Os testes demonstraram que o modelo alcançou uma acurácia de 87,5%, utilizando uma divisão de 70/30% entre os dados de treinamento e teste. O modelo foi capaz de classificar corretamente os alunos em seus respectivos centros com base em padrões predispostos, como a distribuição equilibrada de 50/50 no CCNE, a predominância masculina no CT e a predominância feminina no CE. Esses resultados evidenciam a eficácia da árvore de decisão em capturar essas tendências e fornecer previsões consistentes.