UNIVERSIDADE FEDERAL
DE MINAS GERAIS

CARLOS HENRIQUE BRITO MALTA LEÃO
VINÍCIUS ALVES DE FARIA RESENDE

MINERAÇÃO DE DADOS

ENEM 2022: O Impacto das Características Socioeconômicas no Desempenho dos Candidatos

Belo Horizonte
2023

1. Entendimento do negócio

1.1. Determinar objetivos do negócio

1.1.1. Contexto e relevância

O projeto visa identificar partições nos microdados do ENEM 2022 para analisar como o desempenho dos alunos se relaciona com seus dados pessoais, informações escolares e dados socioeconômicos. Os microdados escolhidos são o nível mais detalhado de dados coletados pelo exame, incluindo informações como provas, gabaritos, detalhes dos itens, notas e respostas do questionário socioeconômico dos participantes.

Nesse contexto, o propósito principal do projeto é aplicar a técnica de agrupamento baseado em representantes (Representative-based clustering). Em outras palavras, o objetivo é encontrar divisões significativas, ou *clusters*, na base de dados, agrupando os participantes de acordo com padrões comuns e características representativas.

A relevância desse projeto é inegável, pois está intrinsecamente relacionada ao direcionamento eficaz de recursos no sistema educacional brasileiro. Ao entender como os fatores pessoais, socioeconômicos e escolares afetam o desempenho dos alunos no ENEM, podemos tomar decisões mais embasadas sobre onde investir recursos educacionais e como direcionar esforços para melhorar a qualidade da educação no país.

1.1.2. Objetivos do negócio

O objetivo pode ser dividido em 4 sub categorias principais, explicitadas a seguir:

Identificação de Padrões de Desempenho: Compreender quais fatores pessoais, socioeconômicos e escolares estão associados ao desempenho no ENEM para orientar programas de apoio.

Equidade na Educação: Garantir que todos os alunos tenham igualdade de oportunidades educacionais, identificando desigualdades no desempenho.

Customização da Educação: Adaptar a educação às necessidades individuais dos alunos para melhorar o aprendizado.

Avaliação da Eficácia Escolar: Avaliar o desempenho das escolas para direcionar recursos adequadamente.

Portanto, o projeto se concentra em fornecer uma análise abrangente dos microdados do ENEM 2022 para atender a esses objetivos do negócio. A análise será realizada com o objetivo de fornecer informações valiosas que orientarão políticas educacionais e estratégias de melhoria no sistema de ensino.

1.1.3. Critérios de sucesso do negócio

Os critérios de sucesso associados a este problema incluem:

Melhorar o Desempenho dos Alunos: O foco é identificar padrões que estejam relacionados ao melhor desempenho dos alunos no ENEM, considerando fatores como dados pessoais, informações escolares e contexto socioeconômico.

Promover a Equidade na Educação: É importante detectar discrepâncias no desempenho entre diferentes grupos demográficos, incluindo grupos étnicos, socioeconômicos e geográficos, a fim de desenvolver políticas que visem à equidade educacional.

Personalização da Educação: Utilizar os resultados para adaptar a abordagem educacional às necessidades individuais dos alunos, visando aprimorar o aprendizado.

Avaliação das Escolas: Além de avaliar o desempenho dos alunos, também buscamos avaliar a eficácia das instituições de ensino, a fim de direcionar apoio e recursos de maneira adequada.

Esses critérios de sucesso serão avaliados ao longo do projeto para garantir que o objetivo seja alcançado de maneira eficaz e eficiente.

1.2. Avaliar a Situação

1.2.1. Inventário de Recursos

Para o projeto de mineração de dados, estará disponível um computador pessoal com as seguintes especificações técnicas:

- CPU: Ryzen 7 5700x
- Memória RAM: 2x16GB a 4800 MHz
- GPU: RTX 2060 SUPER
- Disco de Armazenamento: SSD NVME M.2 de 1TB

Essas especificações proporcionarão um ambiente de processamento robusto e rápido, adequado para a análise de dados e mineração dos agrupamentos nos microdados do ENEM 2022. O hardware estará com plena disponibilidade para o projeto, uma vez que é pertencente a um dos integrantes.

Em relação a fonte dos dados, o Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP), por intermédio da Diretoria de Avaliação da Educação Básica, em cumprimento da sua missão de desenvolver e disseminar informações sobre os exames e avaliações da educação básica, disponibiliza os Microdados do Enem 2022. Os microdados se constituem no menor nível de desagregação de dados recolhidos por pesquisas,

avaliações e exames realizados. No caso do ENEM, os dados estão por participante, de forma anonimizada.

Para atender a demanda dos usuários sobre informações específicas, são disponibilizadas as provas, os gabaritos, as informações sobre os itens, as notas e o questionário socioeconômico respondido pelos inscritos no ENEM. Os dados são disponibilizados em formato ".csv" (formato de arquivo que contém valores separados por delimitador com ponto-e-vírgula) e os inputs para a leitura desses arquivos foram elaborados utilizando os softwares SAS, SPSS e R. Os dados podem ser obtidos de forma gratuita e on-line por meio do portal de Dados Abertos do website gov.br.

No que tange à fonte de conhecimento que será utilizada, é de grande relevância o uso dos dados de censos do IBGE (Instituto Brasileiro de Geografia e Estatística) e derivados. Isso se dá pelo fato de que os dados nortearão fatores socioeconômicos que são relevantes para a análise, além de proporcionar a visão geográfica da distribuição destes indicadores.

Por fim, para obter os resultados almejados, utilizaremos técnicas estudadas durante o curso de Mineração de Dados ofertado pelo DCC (Departamento de Ciência da Computação), principalmente em relação à algoritmos destinados à mineração de *clusters*.

1.2.2. Requisitos, suposições e restrições

Nesta seção, delinearemos os requisitos essenciais, suposições e restrições que guiarão o projeto. Esses elementos são fundamentais para garantir que o projeto seja conduzido de forma eficaz, cumpra seus objetivos e atenda às expectativas.

Requisitos do Projeto:

Cronograma de Conclusão: O projeto deve ser concluído de acordo com o cronograma estabelecido, levando em consideração os prazos definidos para cada fase, desde a obtenção de dados até a análise e apresentação dos resultados em um relatório final. As datas foram definidas pelo professor, sendo a primeira entrega a ser realizada no dia 31 de outubro, a segunda no dia 07 de novembro e a fase de engenharia reversa no dia 14 de novembro.

Compreensão e Qualidade dos Resultados: É fundamental que a análise dos microdados do ENEM 2022 seja realizada com um alto nível de compreensão e qualidade. Isso envolve a aplicação de métodos de mineração de *clusters* adequados e a validação cuidadosa dos resultados obtidos.

Segurança dos Dados: Uma vez que os dados são fornecidos já anonimizados pelo INEP, não há maneira de que dados pessoais sejam vazados, uma vez que a base de dados de referência já não contém nenhum tipo de dado pessoal.

Questões Legais: O projeto deve estar em conformidade com todas as leis e regulamentações aplicáveis. Essa parte torna-se simples uma vez que os dados de referência possuem uma utilização livre definida pelo INEP.

Permissão para Uso dos Dados: Devemos cumprir os termos de uso estabelecidos pelo INEP para garantir o uso ético dos dados.

Premissas do Projeto:

Qualidade dos Dados: Partimos da premissa de que os microdados do ENEM 2022 estão completos e que a qualidade dos dados é adequada para nossa análise. Caso contrário, as limitações dos dados serão consideradas na interpretação dos resultados.

Acesso aos Dados do IBGE: Assumimos que teremos acesso aos dados de censos do IBGE necessários para enriquecer nossa análise com informações socioeconômicas. Caso contrário, nossa capacidade de avaliar o impacto desses fatores pode ser limitada.

Disponibilidade de Recursos Técnicos: Supomos que teremos acesso contínuo aos recursos técnicos necessários para executar as análises de mineração de dados, incluindo hardware e software adequados.

Restrições do Projeto:

Recursos Limitados: Podem existir limitações de recursos humanos e de tempo para realizar todas as tarefas necessárias no projeto. Isso pode afetar a extensão e a complexidade das análises realizadas.

Possibilidade de Viés nos Dados: Reconhecemos que os dados do ENEM 2022 podem conter viés devido a respostas inadequadas ou incorretas no questionário socioeconômico. Isso pode afetar a precisão das conclusões e será levado em consideração na interpretação dos resultados.

É crucial manter esses requisitos, suposições e restrições em mente ao longo do projeto, pois eles influenciarão as decisões tomadas e ajudarão a garantir a realização bem-sucedida da análise de Mineração de Dados do ENEM 2022.

1.3. Determinar os objetivos da mineração de dados

1.3.1. Resultados de mineração de dados

Nesta seção, delinearemos os resultados pretendidos da mineração de dados que nos permitirão alcançar os objetivos de negócios estabelecidos anteriormente. Esses resultados são principalmente técnicos e estão alinhados com as questões sociais que buscamos resolver.

Tradução de Questões Sociais para Objetivos de Mineração de Dados:

Segmentação de Padrões de Desempenho: Traduziremos essa questão definida anteriormente em um objetivo de mineração de dados, que envolve a criação de modelos capazes de segmentar os padrões de desempenho dos alunos no ENEM 2022. Isso implica o agrupamento dos alunos baseados em fatores pessoais, socioeconômicos e escolares que influenciam no seu desempenho.

Equidade na Educação: Nosso objetivo de mineração de dados aqui é desenvolver modelos que identifiquem desigualdades no desempenho educacional entre diferentes grupos demográficos. Isso pode envolver técnicas de *clustering* para entender como as características individuais se relacionam com o desempenho.

Customização da Educação: Para atingir esse objetivo, nossa mineração de dados visa explicitar os padrões de características educacionais com os resultados obtidos. Isso pode ajudar a evidenciar falhas no padrão educacional.

Avaliação da Eficácia Escolar: Neste caso, nosso objetivo é avaliar o desempenho dos alunos em relação ao tipo de instituição de ensino que frequentaram, tanto em nível individual quanto agregado. Utilizaremos técnicas de mineração de *clusters* para entender como as características das escolas estão relacionadas ao desempenho dos alunos.

Tipo de Problema de Mineração de Dados:

Os tipos de problemas de mineração de dados abordados neste projeto incluem principalmente a mineração de *clusters* representativos. Essa técnica será empregada com o objetivo de agrupar os dados de maneira representativa, possibilitando, assim, a análise de desigualdades e a identificação de características similares entre os participantes, de forma a elucidar segmentações de desempenho dadas características comuns.

1.3.2. Critérios de sucesso da mineração de dados

Para determinar o sucesso da mineração de dados neste projeto, definimos os seguintes critérios técnicos:

Desempenho e Complexidade: Avaliaremos o desempenho dos modelos em termos de tempo de execução e recursos computacionais necessários. Procuramos modelos eficientes e escaláveis que possam lidar com grandes volumes de dados.

Explicabilidade do Modelo: A capacidade de explicar as descobertas dos modelos é fundamental. Os resultados devem ser apresentados de forma que sejam compreensíveis e úteis para os tomadores de decisão, permitindo que eles entendam os fatores que influenciam o desempenho dos alunos.

Visão Socioeconômica Fornecida pelo Modelo: Se aplicável, os modelos devem fornecer insights socioeconômicos valiosos, como identificação de grupos-alvo para programas de apoio educacional. Esses insights devem ser relevantes e úteis para as partes interessadas.

É importante notar que esses critérios de sucesso da mineração de dados são diferentes dos critérios de sucesso "negocial" definidos anteriormente. Eles se concentram na qualidade técnica dos modelos e em sua capacidade de fornecer insights valiosos a partir dos dados. Esses critérios serão monitorados e avaliados ao longo do projeto para garantir que os objetivos de negócios sejam alcançados com eficácia.

1.4 - Produzir o Plano de Projeto:

Nesta seção, descreveremos o plano pretendido para alcançar os objetivos de mineração de dados, garantindo que eles estejam alinhados com os objetivos estabelecidos anteriormente. O plano do projeto incluirá as etapas a serem executadas, a duração estimada, os recursos necessários, as entradas e saídas de cada fase, bem como as dependências entre elas. Além disso, serão consideradas as iterações nas fases de modelagem e avaliação, quando apropriado, bem como a análise das dependências entre o cronograma e os riscos identificados.

1.4.1 - Plano do Projeto:

O plano do projeto é composto pelas seguintes etapas, com as respectivas informações detalhadas:

Fase de Preparação de Dados (Duração: 1 dia):

- **Entradas:** Dados brutos do ENEM 2022, critérios de seleção de variáveis.
- **Saídas:** Conjunto de dados preparados e limpos para análise.
- **Dependências:** Essa fase é crítica para todas as etapas subsequentes do projeto.

Fase de Compreensão de Dados (Duração: 1 dia):

- **Entradas:** Conjunto de dados preparados, objetivos abstratos traduzidos em questões de mineração de dados.

- **Saídas:** Relatório de análise exploratória de dados.
- **Dependências:** Depende da conclusão da fase de preparação de dados.

Fase de Modelagem (Duração: 1 dia):

- **Entradas:** Conjunto de dados preparados, relatório de análise exploratória, critérios de sucesso de mineração de dados.
- **Saídas:** Modelos de mineração de dados treinados.
- **Dependências:** Depende da conclusão bem-sucedida da fase de compreensão de dados.

Fase de Implantação (Duração: 2 dias):

- **Entradas:** Modelos de mineração de dados validados.
- **Saídas:** Implantação dos modelos em um ambiente de produção.
- **Dependências:** Depende da conclusão bem-sucedida da fase de avaliação.

Fase de Avaliação dos Resultados e Geração do Relatório Final (Duração: 2 dias):

- **Entradas:** Resultados da avaliação, dados socioeconômicos, dados de desempenho dos candidatos.
- **Saídas:** Relatório final com análise dos resultados e recomendações para políticas educacionais.
- **Dependências:** Depende da conclusão bem-sucedida da fase de implantação.

Pontos de Decisão e Revisão:

Durante cada fase, haverá pontos de decisão para avaliar se os resultados atuais atendem aos critérios de sucesso. Pontos de revisão serão realizados ao final de cada fase para verificar se os resultados estão em conformidade com os objetivos de negócios.

Iterações Importantes:

Iterações nas fases de modelagem e avaliação podem ocorrer, se necessário, para refinar os modelos e garantir que atendam aos critérios de sucesso.

Análise de Dependências entre Cronograma e Riscos:

Será realizada uma análise contínua das dependências entre o cronograma do projeto e os riscos identificados. Ações e recomendações serão implementadas conforme necessário para mitigar riscos que possam afetar o andamento do projeto.

1.4.2 - Avaliação Inicial de Ferramentas e Técnicas:

No final da primeira fase, a equipe do projeto realizará uma avaliação inicial de ferramentas, técnicas e algoritmos para determinar as mais adequadas ao projeto. Isso incluirá a seleção de uma ferramenta de mineração de dados que suporte métodos apropriados para todas as fases do processo.

Para a realização do projeto, temos algumas atividade definidas. Primeiramente, é necessário a criação de critérios de seleção para ferramentas e técnicas. Além disso, a escolha de potenciais algoritmos para a mineração de *clusters* também é importante. Por fim, também faz-se necessário a avaliação da adequação dos algoritmos em relação aos objetivos já definidos.

Este plano de projeto será consultado continuamente e revisado ao longo do projeto, garantindo que as atividades estejam alinhadas com os objetivos comerciais e que as dependências e riscos sejam gerenciados adequadamente.

2. Entendimento dos dados

2.1. Coletar dados iniciais

2.1.1. Relatório inicial de produção de dados

Planejamento de requisitos de dados

Para este projeto, a ideia é realizar paralelos entre indicativos socioeconômicos e as performances dos candidatos. Dessa forma, precisaremos de informações que possam indicar o desempenho do aluno, tanto quanto a performance do candidato que realizou a prova tanto do candidato que foi desistente. Ademais, será necessário que métricas socioeconômicas bem definidas estejam disponíveis, para que padrões relacionando desempenho e características socioeconômicas sejam encontradas.

Após analisar o Dicionário de Microdados do Enem 2022 também disponibilizado pelo INEP, conseguimos verificar que as informações necessárias para o projeto estão disponíveis em plenitude.

Critério de seleção

Para alcançar os objetivos definidos neste projeto, nem todos os atributos da base de dados são necessários. Portanto, determinamos os atributos mais relevantes para atingir o objetivo proposto.

Em primeiro lugar, selecionamos alguns atributos relacionados à situação socioeconômica do participante. No que diz respeito aos dados do participante, iremos analisar

a faixa etária, gênero, estado civil, cor/raça e o tipo de escola do Ensino Médio (pública ou privada). Além disso, também utilizaremos as respostas aos vinte e cinco itens do questionário socioeconômico preenchido pelo participante durante a inscrição.

Em segundo lugar, também será necessário examinar os atributos relacionados aos resultados obtidos pelo participante. Nesse contexto, realizaremos uma análise dos dados da prova objetiva, incluindo a presença do participante, bem como sua pontuação em cada uma das quatro provas (Ciências da Natureza, Linguagens e Códigos, Matemática). Além disso, também consideraremos os dados relacionados à redação, incluindo a situação da redação do participante e sua respectiva nota.

Por fim, também será imprescindível a exclusão dos dados associados aos candidatos que realizaram a prova exclusivamente para fins de treinamento de seus conhecimentos. Além disso, será essencial eliminar quaisquer observações que contenham atributos ausentes, uma vez que a presença de valores nulos pode comprometer a integridade e a confiabilidade da análise de dados. Portanto, a identificação e exclusão de objetos de dados com atributos faltantes são medidas importantes para assegurar a qualidade dos dados e a eficácia da análise subsequente.

Dessa forma, ao focar em atributos específicos relacionados à situação socioeconômica e aos resultados dos participantes, poderemos alcançar de maneira mais eficaz os objetivos estabelecidos para este projeto.

Inserção de dados

Para o projeto em questão e os dados que deseja-se minerar, apenas dois dados não estão presentes e terão de ser inseridos. Um deles é o que se refere à média da nota da prova, ou seja, uma forma de sumarizar todas as competências da prova do ENEM em uma única métrica. A ideia para a obtenção deste dado é simplesmente a soma de todas as notas e a divisão por cinco que equivale ao número de avaliações distintas, tal metodologia é amplamente utilizada, inclusive para a seleção via SISU da Universidade Federal de Minas Gerais (UFMG).

Outro dado que temos a intenção de conseguir é um classificador categórico para os intervalos das notas dos candidatos, ou seja, ao invés de procurar por padrões onde fatores socioeconômicos apontam para uma nota específica, o objetivo é encontrar esses indicadores socioeconômicos que apontam para um intervalo de notas. Dessa forma, poderemos classificar a nota do candidato de forma categórica, o que viabilizará o processo de mineração de *clusters*.

Referências Bibliográficas

INSTITUTO NACIONAL DE ESTUDOS E PESQUISAS EDUCACIONAIS ANÍSIO TEIXEIRA. **Microdados do Enem 2022.** Brasília: Inep, 2023. Disponível em: <https://www.gov.br/inep/pt-br/acesso-a-informacao/dados-abertos/microdados/enem>. Acesso em: 19 abr. 2023.