**Proposta Trabalho Final - Pós Graduação Senac Rio**

**Especialização Big Data**

1. **Nome dos integrantes do grupo:**

Bruno Carvalho

Eduardo Coutinho

Ricardo Reis

1. **Título provisório do trabalho:**

INDICADORES SÓCIOECONÔMICOS DOS CANDIDATOS DO ENEM ATRAVÉS DE MICRODADOS DISPONIBILIZADOS PELO PORTAL INEP

ANÁLISE SÓCIOECONÔMICA DOS CANDIDATOS DO ENEM ATRAVÉS DE MICRODADOS DISPONIBILIZADOS PELO PORTAL INEP

1. **Área ou disciplinas ou assuntos tratados no trabalho (vide item 2 acima):**

ÁREA DA EDUCAÇÃO

1. **Descrição de aplicabilidade (pode ser parecido, extraído ou inspirado no título):**
   1. A aplicabilidade do projeto se deve ao estudo dos dados disponibilizadas pelo INEP, para a realização do cruzamento das informações.
   2. Após análise dos microdados do ENEM, iremos buscar a correlação das notas de aprovação com relação fatores socioeconômicos do candidato.
   3. Após essas verificações, poderemos validar o quão importante é para o aluno de ensino médio estar com uma boa infraestrutura de ensino, para ter um resultado satisfatório no ingresso a uma universidade no Brasil.
2. **Contexto do trabalho (qual o contexto da empresa, como descreveria como é o mundo real ou fictício no qual o projeto está sendo desenvolvido?):**
   1. O contexto do trabalho é identificar os principais problemas da educação no Brasil, através da análise dos microdados disponibilizados pelo INEP, com o objetivo de nivelar a educação do ensino médio para acesso às universidades em todo o território nacional.

A extração dos dados educacionais tem se tornado um campo em amplo desenvolvimento para estudos e obtenção de padrões pedagógicos em diversas instituições públicas e particulares. Através dessa análise é possível um reconhecimento no processo de aprendizado em bases da educação de nível básico à nível médio.

O Exame Nacional do Ensino Médio (Enem), foi criada em 1998. Ela é utilizada para avaliar a qualidade do ensino médio no país. Seu resultado serve para acesso ao ensino superior em universidades públicas brasileiras, através do Sistema de Seleção Unificada (SiSU), assim como em algumas universidades no exterior. O Enem é o maior exame vestibular do Brasil e o segundo maior do mundo, atrás somente do Gāo Kǎo, o exame de admissão do ensino superior da República Popular da China.

É uma prova realizada pelo Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP), sendo uma autarquia federal vinculada ao Ministério da Educação (MEC). Seu objetivo é promover estudos, pesquisas e avaliações periódicas sobre o sistema educacional brasileiro, com o objetivo de subsidiar a formulação e implementação de políticas públicas para a área educacional.

O INEP realiza levantamentos estatísticos e avaliações em todos os níveis e modalidades de ensino. Suas atividades mais conhecidas são: a realização do Exame Nacional do Ensino Médio (Enem); a realização do Exame Nacional de Desempenho de Estudantes (Enade); a organização das avaliações periódicas dos ensinos básico e superior brasileiro.

Diversas universidades do Brasil, utilizam-se como critério de admissão a nota de qualificação obtida na prova e posteriormente é possível realizar o download de um *dataset* disponibilizado referente a cada ano de prova feita pelos alunos.

Neste estudo realizaremos uma verificação sobre a possibilidade de previsão das notas dos alunos inscritos, aplicando somente a categoria socioeconômica atribuída ao questionário no momento da inscrição para realização da prova.

Abordaremos diversos temas relacionados ao assunto de big data, dentre eles : coleta de dados, através dos *datasets* disponibilizados pelo INEP, limpeza dos dados, onde aplicaremos uma correlação entre tabelas e dicionários com informações referentes as bases de dados disponíveis, pré-processamento, onde faremos a eliminação de anomalias que de alguma forma possa comprometer por total o resultado e qualidade de todo processo, mineração dos dados, utilizando Apache Spark (ferramenta Big Data que tem o objetivo de processar grandes conjuntos de dados de forma paralela e distribuída) e Jupyter Notebook para a verificação dos dados já totalmente validados e sólidos, onde será feita a conexão à um banco de dados MySQL. A automatização de todos esses processos, acarretam uma diminuição em possibilidades futuras de erro pela interferência humana, onde questões que poderiam passar totalmente despercebidas são identificadas por essa análise. Em suma, a visualização de informações é um processo extremamente intuitiva para apresentação dos dados que foram estudados e analisados e com isso para o sucesso de todo o projeto. Todas as informações foram colhidas, analisadas, filtradas e expostas para um fácil entendimento dos dados, para isso adaptações formam um fácil entendimento sobre todo o assunto abordado, através de gráficos que disponibilizaremos com geolocalização feita através do matplotlib, uma biblioteca de software para criação de gráficos e visualizações de dados em geral, feita para a da linguagem de programação Python e sua extensão de matemática *NumPy*, onde esse processo será descrito na ferramenta Jupyter Notebook conforme citado anteriormente.

1. **Problema a ser solucionado (foco no problema, de forma detalhada):**

O presente estudo aplica-se a verificação e validação de *datasets* provenientes do Exame Nacional do Ensino Médio – ENEM, através do qual será feita uma análise exploratória dos dados, possibilitando a identificação de precariedade de estudo em determinadas regiões do Brasil e relacioná-las.

O objetivo inicial é aferir se:

1. Os candidatos estão em condições de desenvolver todos os seus conhecimentos científicos e tecnológicos;
2. Identificar a geolocalização de instituições de ensino médio e correlacionar com as notas providas da base de dados do INEP;
3. Apresentar a ordenação de resultados por áreas de aplicação, dentre elas: Matemática, Ciências Humanas, Ciências da Natureza, Linguagens e Códigos e Redação;
4. Apresentar correspondência com a infraestrutura familiar na interferência direta e indireta do resultado dos alunos ao acesso às Universidades.

6.1 Estudo do *Dataset*:

O exame passou por alterações estruturais entre os exercícios de 2014 e 2015, algumas dessas refletindo na análise dos microdados.

O projeto baseia-se em uma base de dados, que contém em torno 20 gigabytes, utilizando *datasets* entre os anos de 2015 a 2019, provida pelo INEP (Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira), onde a estrutura dos dados é estruturada e gerada através de arquivos com extensão (.csv). Um arquivo CSV (Valores Separados por Vírgula) é um tipo especial de arquivo que você pode criar ou editar no Excel. Em vez de armazenar informações em colunas, os arquivos CSV armazenam informações separadas por vírgulas.

Em vez de armazenar informações em colunas, os arquivos CSV armazenam informações separadas por vírgulas. sendo o processo de estudo dividido em etapas de atuação como descrito abaixo:

* + 1. A primeira etapa da análise do *dataset*, consiste na verificação dos dados que o candidato preenche no formulário de inscrição para a prova, com isso, obtêm-se diversos campos de informação, como: município de residência, cidade, nacionalidade, município de nascimento, idade, cor/raça, sexo, estado civil, grau de escolaridade do pai e da mãe, tipo de escola oriunda (instituição pública ou privada) e se o estudante apresenta algum tipo de deficiência.

6.1.2 Segunda etapa da análise do *dataset*, apresentam-se dados referentes à questões socioeconômicas do candidato, realizadas no ato da inscrição do exame. Abaixo apresentamos campos que são preenchidos:

* + 1. Até que série seu pai, ou o homem responsável por você, estudou?
    2. Até que série sua mãe, ou a mulher responsável por você, estudou?
    3. Incluindo você, quantas pessoas moram atualmente em sua residência?
    4. Qual é a renda mensal de sua família? (Some a sua renda com a dos seus familiares.)
    5. Na sua residência tem telefone celular?
    6. Na sua residência tem computador?
    7. Na sua residência tem acesso à Internet?
    8. Você já concluiu ou está concluindo o Ensino Médio?
    9. Em que tipo de escola você frequentou o Ensino Médio?

6.1.3 Terceira etapa da análise do *dataset*, consiste na exposição ao rendimento dos alunos nas provas objetivas e redação. Com isso, é realizado um estudo nos seguintes campos abaixo:

1. Presença na prova objetiva de Ciências da Natureza
2. Presença na prova objetiva de Ciências Humanas
3. Presença na prova objetiva de Linguagens e Códigos
4. Presença na prova objetiva de Matemática
5. Nota da prova de Ciências da Natureza
6. Nota da prova de Ciências Humanas
7. Nota da prova de Linguagens e Códigos
8. Nota da prova de Matemática
9. Língua Estrangeira
10. Situação da redação do participante
11. Nota da prova de redação

6.2 Arquitetura de Dados:

Com isso, para uma melhor análise e visualização na aplicabilidade aos termos e técnicas de todo processo do nosso estudo, foi desenvolvido uma arquitetura simplificada, vide (Fig.1) objetivando todas as ferramentas e procedimentos que serão aplicados e abordados em todo projeto.

Texto preto sobre fundo branco

Descrição gerada automaticamente

Fig.1 (Fonte: Elaboração dos Autores)

6.3. Mineração de dados:

A mineração de dados é dita como um processo utilizado para extração de dados de um conjunto maior de dados brutos e se tem a aplicação em diversos campos como: ciência e pesquisa. Envolve coleta, armazenamento de dados eficazes, pré-processamento, transformação dos dados (limpeza), busca de padrões e visualização final dos dados, gerando o conhecimento.

Abaixo temos (Fig.2), que apresenta todo o procedimento da fase de início ao fim do método aplicado.

Uma imagem contendo texto, mapa

Descrição gerada automaticamente

Fig.2

Estágios do Processo KDD e Data Mining.

*FONTE: Fayyad, Piatetsky-Shapiro e Smyth.****From data mining to knowledge discovery in databases.***

6.4 Pré-Processamento:

A fase de pré-processamento talvez seja a mais importante a fim de se obter um bom resultado, o pré-processamento nada mais é do que o processo de preparação, organização e estruturação dos nossos dados, além de ser o momento ideal para escolhermos quais dados fazem sentido estarem sendo analisados em nosso *dataset*.

É importante realçar a importância desta fase, dito que, a qualidade de dados pode influenciar diretamente no resultado na modelagem, muitas das vezes pensamos que o problema da nossa solução é o algoritmo usado para gerar o modelo, porém o grande vilão é o seu próprio conjunto de dados que podem possuir muitos atributos com valores faltantes, outliers e escalas de valores contradizentes e por fim nenhum modelo será capaz de trabalhar com esses dados e gerar resultados de qualidade.

Para a análise de estudo aos microdados disponibilizados pelo INEP (Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira) em relação ao ENEM, teve-se a aplicação das seguintes ferramentas:

* Apache Spark
* Jupyter Notebook
* Banco de Dados – MySQL
* Visual Studio Code
* Microsoft PowerBI

Para analisar e validar as bases de dados foram aplicadas de forma local (máquina física), onde tivemos alguns problemas de processamento de arquivos do *dataset*, por possuírem uma quantidade excessiva de informações. Para isso, feita uma análise e retiradas colunas que não seriam de suma importância para a abrangência do estudo, com isso restando 49 colunas onde faremos uma abordagem socioeconômica de acordo com itens do item 6.1. Anteriormente havíamos pensado em estar exportando para uma *engine* de banco de dados em Cloud, porém ficaria totalmente inapropriado em ter um custo muito elevado para aplicação desses dados dessa forma. Com isso decidimos aplicar em um desktop robusto e com configurações que iriam suportar todo processamento desses microdados.

6.4.1 – Apache Spark

O Apache Spark é uma ferramenta Big Data que tem o objetivo de processar grandes conjuntos de dados de forma paralela e distribuída. Visto que a massa de dados utilizada é de grande quantidade vide item 6.1, com isso a arquitetura utilizada pelo Spark em nosso projeto, conta com verificação do conjunto de informações e dados disponibilizados e em paralelo sendo feito o desenvolvimento da aplicação com o Banco de Dados Mysql e linguagem de programação Python.

6.4.2 – Jupyter Notebook

O Jupyter Notebook é uma ferramenta de computação alfabetizada, extremamente eficiente, pois com ela é possível unir códigos e textos. Assim é possível explicar cada funcionalidade de forma detalhada. É possível gerar gráficos em tempo real dentro da ferramenta. Em nosso estudo foi desenvolvido para desenvolvimento e documentação dos códigos de forma inicial, gerando um grande fluxo e rapidez na visualização do resultado da análise do *dataset*.

6.4.3 – Banco de Dados Mysql

O MySQL é um sistema de gerenciamento de banco de dados (SGBD), que utiliza a linguagem SQL (Linguagem de Consulta Estruturada, do inglês *Structured* Query *Language*) como interface. É atualmente um dos sistemas de gerenciamento de bancos de dados mais populares da Oracle Corporation, com mais de 10 milhões de instalações pelo mundo.

Após análise e limpeza dos dados, foram exportadas as colunas para uma base de dados (Mysql), onde será feita a conexão com uma ferramenta de visualização para exposição dos dados gerados por *querys* no banco de dados.

6.4.4 – Visual Studio Code

O Visual Studio Code é um editor de código-fonte desenvolvido pela Microsoft para Windows, Linux e macOS. Ele é um software livre e de código aberto, apesar do download oficial estar sob uma licença proprietária.

Após utilização inicial do Jupyter Notebook, resolvemos fazer a migração para o Visual Studio *Code* e com isso tivemos uma performance no processamento. Ferramenta onde desenvolvemos o código-fonte do projeto e toda documentação linha a linha.

6.4.4 – Visualização de Dados com PowerBI

O PowerBI é um serviço de Business Analytics baseado em nuvem que oferece um ponto de visão única dos dados comerciais mais críticos. É possível também realizar um monitoramento da saúde do negócio da empresa usando Dashboards interativos e em tempo real, criar relatórios e muito mais.

Com isso no nosso projeto, a parte visualização dos dados finais foi abordada com o software Microsoft PowerBI, prestando uma automação e sincronização com a base de dados Mysql, retornando com uma performance praticamente instantânea e mantendo a integridade dos dados após realização de procedimentos internos. Foi possível a geração de relatórios quantitativos e qualitativos através da ferramenta e apresentação de Dashboards com aplicação em Cloud e na Web para visualização em Real Time.

6.5 - Data Cleaning

Em virtude da análise feita no *dataset*, foram encontrados dados de diversos estudantes que faltaram à prova, campos não preenchidos de forma correta, onde se tornou necessária a remoção dessas informações e as que não se fazem referência ao nosso projeto.

Para tal procedimento de limpeza dos *datasets*, foi feita a seleção das colunas através do dicionário disponibilizado em conjunto com a base de dados, onde a documentação estrutural da está contida e com isso retiramos as que seriam úteis para o nosso projeto e que se fazem correlacionais com a questão socioeconômica do candidato.

Toda essa transformação e alteração dos dados, foram feitas através de *querys* na aplicação do banco de dados estruturado Mysql e com isso, retornando informações concretas após longo tempo de processamento na base de dados.

6.6 – Análise Exploratória

Em suma a análise exploratória de dados é considerada um dos pontos mais cruciais para o processo de análise em bases de dados. Quer dizer que é um processo que adota: organização, resumo, aplicação de alguns ou vários cálculos e visualização dos dados presentes na análise. Esse processo foi desenvolvido pelo grande estatístico, *John W. Tukey* e com o livro *Exploratory Data* *Analysis (EDA)* lançado em 1977, fez uma contribuição memorável para o mundo científico e social, unindo pensamento estatístico aos processos de transformação e exploração de dados.

Com isso em nossa análise exploratória está sendo feita a aplicação em *datasets* providos do Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira - INEP e realizando a verificação para a possibilidade de diminuirmos a motivo da falta de acesso ao ensino superior derivado de questões socioeconômicas.

Para isso adotamos em nosso projeto, aplicações em linguagem Python, desenvolvendo os códigos com integração a *engine* de banco de dados relacional e com isso possamos compreender, gerar tabelas de dados, visualizações concretas, exposição de fatos numéricos com base em fórmulas e modelos estatísticos, para enfim podermos apresentar a visualização dos dados.

Em torno do processo abordaremos a identificação dos padrões de cada item, saber quais seus pontos fortes e fracos, diminuir a desigualdade e com isso buscando uma padronização final na análise dos *datasets* colhidos através da base de dados.

6.6.1 – Regressão Linear

Sendo uma grande aliada para o acompanhamento de possíveis cenários de um negócio e suporte para uma tomada de decisão. Uma aplicação da regressão linear se deve em entender um determinando conjunto de variáveis independentes como por exemplo: sexo, raça, nacionalidade, estado civil, idade, oriundos do nosso projeto e está estatisticamente relacionado com o exame e com isso gerando uma amostra de resultado. A nossa abordagem realiza a análise de regressão para ver o quão relacionada é cada variável e isso geralmente requer uma interação para encontrar uma possível combinação certa de variáveis para aplicação a um melhor modelo.

Com a técnica de regressão linear realizaremos um apontamento de quais regiões do país apresentam maior defasagem socioeconômica em relação a questões familiares, de falta de acesso à informação ou uma base social para uma qualidade de ensino. Com isso aplicamos códigos de integração com os *datasets* apontando colunas que fazem referência direta a níveis de estudo dos familiares de cada candidato.

6.7- Visualização dos Dados

Para a etapa de visualização dos dados, abordaremos a ferramenta Microsoft PowerBI onde será feita integração com *engine* de banco de dados Mysql, que será feita a importação dos dados e serão visualizados através de Dashboards dinâmicos , relatórios interativos, gráficos ou diagramas que serão abordados assuntos como : qual região do Brasil com maior índice de notas baixas na área de Ciências da Natureza ou até mesmo em qualquer área abordada pelo exame.

Temos uma melhor diagramação e apresentação do estudo, para que os insights gerados possam promover cada vez mais a tomada de decisão em algum órgão estadual ou federal, dentro do escopo de nossa atuação do projeto. Após a verificação feita na etapa de análise exploratória dos dados é possível apresentar padrões ou conexões com maior rapidez e eficiência.

1. Proposta do rascunho da solução (foco na solução, pelo menos um rascunho sem aprofundamento da mesma)
   1. Através da análise dos microdados disponibilizados pelo INEP, apresentaremos soluções para os problemas que foram abordados no item 6 da proposta do trabalho final.

O presente estudo mostrou que realmente há uma degradação no acesso ao ensino superior em referência às escolas públicas e particulares no país, decorrente de critérios socioeconômicos e oriunda do nível de instrução educacional familiar e de critérios como acesso à tecnologia, dificultando um maior envolvimento com a disseminação de conhecimento e informações.

O projeto apresenta também uma aplicação de limpeza e tratamento dos dados utilizados inicialmente. Posteriormente, com as informações geradas através de ferramentas de Big Data e as correlacionando entre *datasets*, chegamos a resultados estatísticos de alta relevância no índice da educação brasileira e de total prospecção e base para análises futuras de outros exames do Exame Nacional do Ensino Médio (ENEM).

Deve-se, também, ao fato de que a intensificação do repasse de ensino aos familiares de cada estudante que futuramente vá se candidatar ao exame é de extrema importância, pois o apoio e apontamento do jovem em relação a este âmbito é totalmente espelhado.

A solução principal seria um investimento em políticas educacionais e sociais para que se possa delinear estratégias de aprimoramento no ensino básico de acesso e intensificar a disseminação pelo repasse de informações em paralelo com a tecnologia.

Necessita-se de um aprimoramento de trabalho entre o governo federal e o estadual de cada região, para que as instituições de ensino público e particular de todo país tenham uma melhoria expressiva na qualidade de repasse em seus ensinos e seus parques tecnológicos.

1. **Referências:**

http://inep.gov.br/microdados

https://g1.globo.com/educacao/enem/2019/noticia/2019/11/10/enem-2019-tem-menor-taxa-de-abstencao-da-historia-diz-ministerio.ghtm

<https://www.scielo.br/scielo.php?pid=S2176-66812013000200005&script=sci_arttext>

<http://td.inep.gov.br/ojs3/index.php/td/issue/view/390/31>

https://sol.sbc.org.br/index.php/ctic/article/view/3244/3206

http://www.schwartzman.org.br/simon/desafios/4ensinomedio.pdf

https://repositorio.unb.br/handle/10482/20129

https://pt.wikipedia.org/wiki/Instituto\_Nacional\_de\_Estudos\_e\_Pesquisas\_Educacionais\_An%C3%ADsio\_Teixeira

https://blog.leucotron.com.br/etapas-de-big-data/

https://pt.wikipedia.org/wiki/Matplotlib

https://www.devmedia.com.br/introducao-ao-apache-spark/34178

https://www.scielo.br/pdf/rbef/2015nahead/1806-1117-rbef-S1806-11173710001.pdf

https://www.redalyc.org/pdf/2750/275043450074.pdf

https://itforum365.com.br/colunas/kdd-e-data-mining-mais-do-que-apenas-conceitos/

https://www.datageeks.com.br/pre-processamento-de-dados/

https://medium.com/@caiquecoelho/um-guia-completo-para-o-pr%C3%A9-processamento-de-dados-em-machine-learning-f860fbadabe1

https://pt.wikipedia.org/wiki/MySQL

https://pt.wikipedia.org/wiki/Visual\_Studio\_Code

https://medium.com/@suzana.svm/por-que-usar-jupyter-notebook-77d5a59b42a1

http://www.abrapecnet.org.br/enpec/x-enpec/anais2015/resumos/R1960-1.PDF

https://www.revistaespacios.com/a17v38n05/a17v38n05p12.pdf

https://www.datascienceacademy.com.br/course?courseid=microsoft-power-bi-para-data-science

https://medium.com/@gabriel.stankevix/analise-explorat%C3%B3ria-de-dados-732007ddbfaf

https://medium.com/@ligiagalvao/regress%C3%A3o-linear-um-dos-tipos-mais-importantes-de-an%C3%A1lise-de-dados-827cfdbd2229