UFMG
UNIVERSIDADE FEDERAL
DE MINAS GERAIS

CARLOS HENRIQUE BRITO MALTA LEÃO
VINÍCIUS ALVES DE FARIA RESENDE

MINERAÇÃO DE DADOS

Avaliação Preditiva do Desempenho Acadêmico: Mineração de Classificadores para Identificação de Fatores de Evasão e Permanência Estudantil

Belo Horizonte
2023

1. Entendimento do negócio

O sucesso de qualquer projeto de Mineração de Dados reside na compreensão aprofundada dos objetivos e contextos específicos do negócio. No caso desta análise, o foco recai sobre a identificação de fatores cruciais que influenciam a evasão de estudantes e o alcance do sucesso acadêmico em uma instituição de ensino superior. Este processo de entendimento do negócio é vital para garantir que as estratégias e insights resultantes se alinhem de maneira precisa com as metas e necessidades da instituição.

Através de uma análise holística que abrange desde a demografia estudantil até as condições econômicas regionais, buscamos criar um modelo preditivo robusto que não apenas antecipe a evasão, mas também forneça informações acionáveis para melhorar a retenção e o desempenho acadêmico.

1.1. Determinar objetivos do negócio

A etapa inicial deste projeto consistiu na determinação cuidadosa dos objetivos de negócios, uma fase crucial para orientar todo o desenvolvimento subsequente. O principal objetivo do presente projeto é desenvolver um modelo preditivo que possa discernir os fatores determinantes para a evasão estudantil e o sucesso acadêmico.

1.1.1. Contexto e relevância

No início deste projeto de Mineração de Dados, foi realizada uma análise abrangente para coletar informações sobre a situação de negócios de instituições de ensino superior. Esses detalhes foram essenciais para identificar de perto os objetivos de negócios a serem alcançados e para reconhecer os recursos, tanto humanos quanto materiais, que podem ser utilizados ou necessários durante o curso do projeto.

Durante essa fase, foram identificados os seguintes pontos-chave:

- **Demografia Estudantil:** Compreensão detalhada da diversidade dos alunos, incluindo informações demográficas como estado civil, nacionalidade, e status internacional.

- **Condições Socioeconômicas:** Análise das condições socioeconômicas dos alunos, incluindo informações sobre deslocamento, necessidades educacionais especiais, endividamento e situação de bolsa de estudos.
- **Desempenho Acadêmico:** Avaliação dos resultados acadêmicos dos alunos, incluindo o número de unidades curriculares creditadas, matriculadas, avaliadas e aprovadas no primeiro semestre.
- **Informações Econômicas da Região:** Coleta de dados sobre a taxa de desemprego, taxa de inflação e Produto Interno Bruto (PIB) da região, a fim de compreender como esses fatores econômicos podem influenciar as taxas de evasão e o sucesso acadêmico.

Essas informações fornecem um contexto valioso para a análise de dados e orientarão o desenvolvimento de estratégias para prever a evasão estudantil e promover o sucesso acadêmico.

1.1.2. Objetivos do negócio

O principal objetivo do projeto é desenvolver um modelo preditivo eficaz que possa identificar os fatores determinantes para a evasão de estudantes e o sucesso acadêmico. Esse modelo será instrumental para entender os motivos pelos quais os alunos podem abandonar seus estudos ou alcançar sucesso acadêmico, permitindo a implementação de estratégias proativas de retenção e suporte.

Além do objetivo principal, há vários objetivos comerciais relacionados que o cliente gostaria de abordar, tais como:

- Identificar fatores demográficos associados ao sucesso acadêmico.
- Avaliar o impacto das condições socioeconômicas na evasão estudantil.
- Compreender a relação entre o desempenho acadêmico inicial e o sucesso futuro.
- Analisar como as condições econômicas regionais podem influenciar os resultados acadêmicos.
- Esses objetivos adicionais contribuirão para uma compreensão mais holística dos fatores que afetam os resultados dos alunos e fornecerão insights valiosos para a tomada de decisões estratégicas.

1.1.3. Critérios de sucesso do negócio

Os critérios para um resultado bem-sucedido neste projeto foram definidos em termos mensuráveis e específicos para o negócio. Os critérios incluem:

- **Precisão do Modelo:** O modelo deve atingir uma precisão de previsão que exceda um limite predeterminado, garantindo confiança nos resultados obtidos.
- **Identificação Eficaz de Fatores Preditivos:** O modelo deve identificar e quantificar de maneira eficaz os fatores que mais contribuem para a evasão estudantil e o sucesso acadêmico.
- **Interpretabilidade do Modelo:** O modelo deve ser interpretável, permitindo que os tomadores de decisão compreendam facilmente as relações entre variáveis e tomem ações informadas.
- **Aplicabilidade Prática:** As recomendações derivadas do modelo devem ser práticas e acionáveis, proporcionando uma base sólida para a implementação de estratégias de intervenção.

Estes critérios garantirão que o projeto atenda às expectativas do cliente e proporcione valor tangível à instituição de ensino superior na promoção do sucesso acadêmico e na redução da evasão estudantil.

1.2. Avaliar a Situação

A etapa de avaliação da situação é crucial para garantir que todos os recursos, restrições, pressupostos e outros fatores relevantes sejam devidamente considerados no desenvolvimento do plano do projeto de Mineração de Dados.

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 de classificadores. O hardware estará com plena disponibilidade para o projeto, uma vez que é pertencente a um dos integrantes.

O conjunto de dados utilizado neste projeto, intitulado "Predict students' dropout and academic success," foi desenvolvido pelos pesquisadores Valentim Realinho, Jorge Machado, Luís Baptista, e Mónica V. Martins. Este conjunto de dados foi adquirido de diversas bases de dados desconexas de uma instituição de ensino superior e concentra-se em alunos

matriculados em diferentes cursos de graduação, como agronomia, design, educação, enfermagem, jornalismo, administração, serviço social e tecnologias.

A descrição do conjunto de dados abrange informações disponíveis no momento da matrícula dos alunos, incluindo seu percurso acadêmico, dados demográficos e fatores socioeconômicos. Além disso, são registrados o desempenho acadêmico dos alunos ao final do primeiro e segundo semestres.

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 obtenção de um modelo de classificação.

1.2.2. Requisitos, suposições e restrições

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 21 de novembro, a segunda no dia 23 de novembro e a fase de engenharia reversa com data de finalização ainda não definida.
- **Compreensão e Qualidade dos Resultados:** A análise de dados deve resultar em conclusões compreensíveis e de alta qualidade para orientar as decisões estratégicas. Isso envolve a aplicação de métodos de mineração de classificadores adequados e a validação cuidadosa dos resultados obtidos.
- **Segurança dos Dados:** Uma vez que os dados são fornecidos já anonimizados, 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:** No que diz respeito às questões legais associadas a este conjunto de dados, é importante observar que a pessoa que associou este trabalho a este documento dedicou-o ao domínio público, renunciando a todos os seus direitos sobre o trabalho em todo o mundo, de acordo com as leis de direitos autorais, incluindo todos os direitos relacionados e vizinhos, na medida permitida por lei.

Pressupostos do Projeto:

- **Sobre os Dados:** Assume-se que os dados disponíveis são representativos, suficientes e verdadeiros para realizar uma análise significativa.

- **Sobre o Projeto:** Pressupõe-se que a implementação das estratégias derivadas da análise resultará em melhorias mensuráveis, tanto academicamente quanto socialmente.

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 encontrados podem conter viés devido a respostas inadequadas ou incorretas no processo de matrícula dos estudantes, assim como um armazenamento incorreto destas informações. Isso pode afetar a precisão das conclusões e será levado em consideração na interpretação dos resultados.

Ao listar claramente esses requisitos, pressupostos e restrições, garantimos transparência e alinhamento com as expectativas do cliente, mitigando possíveis desafios ao longo do projeto.

1.3. Determinar os objetivos da mineração de dados

1.3.1. Resultados de mineração de dados

O principal objetivo de negócio para este projeto é desenvolver um modelo preditivo capaz de identificar os fatores determinantes para a evasão estudantil e o sucesso acadêmico. Traduzindo este objetivo para os termos técnicos da mineração de dados, os resultados pretendidos incluem:

- **Modelo de Classificação:** Desenvolver um modelo de classificação que seja capaz de categorizar os alunos em dois grupos distintos ao final da duração normal do curso: evasão ou matriculado/graduado.
- **Identificação de Fatores Preditivos:** O modelo deve ser capaz de identificar e quantificar eficazmente os fatores que mais contribuem para a evasão estudantil e o sucesso acadêmico. Esses fatores podem incluir características demográficas, condições socioeconômicas, desempenho acadêmico inicial e outros relevantes.
- **Interpretabilidade do Modelo:** O modelo deve ser interpretável, permitindo que os tomadores de decisão compreendam as relações entre as variáveis e possam tomar ações informadas com base nos insights fornecidos.
- **Acurácia Preditiva:** Alcançar uma precisão preditiva satisfatória que exceda um limite predeterminado, garantindo confiança nos resultados obtidos.

1.3.2. Critérios de sucesso da mineração de dados

Os critérios de sucesso da mineração de dados são fundamentais para avaliar a eficácia do modelo e sua contribuição para os objetivos de negócios. Definimos os seguintes critérios para um resultado bem-sucedido em termos técnicos:

- **Precisão do Modelo:** O modelo deve atingir uma precisão de previsão que exceda um limite predeterminado, garantindo confiança nos resultados obtidos.
- **Identificação Eficaz de Fatores Preditivos:** O modelo deve identificar e quantificar de maneira eficaz os fatores que mais contribuem para a evasão estudantil e o sucesso acadêmico.
- **Interpretabilidade do Modelo:** O modelo deve ser interpretável, permitindo que os tomadores de decisão compreendam facilmente as relações entre variáveis e tomem ações informadas.
- **Aplicabilidade Prática:** As recomendações derivadas do modelo devem ser práticas e acionáveis, proporcionando uma base sólida para a implementação de estratégias de intervenção.
- **Benchmark de Avaliação:** Estabelecer benchmarks para avaliação dos critérios, comparando o desempenho do modelo com padrões de referência e melhores práticas reconhecidas.
- **Critérios Subjetivos:** Além dos critérios objetivos, considerar critérios subjetivos, como a capacidade de explicação do modelo e a contribuição para a visão de marketing fornecida pelo modelo.

Ao definir esses critérios de sucesso, garantimos que a avaliação do modelo seja abrangente e alinhada com os objetivos de negócios, proporcionando uma base sólida para as decisões estratégicas baseadas nos resultados da mineração de dados.

1.4 - Produzir o Plano de Projeto:

O sucesso da mineração de dados depende de um plano de projeto abrangente que guie todas as atividades, desde a coleta de dados até a implementação das soluções propostas. Este plano é dinâmico e será consultado continuamente e revisado ao longo do projeto, especialmente ao iniciar novas tarefas ou iterações.

1.4.1 - Plano do Projeto:

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

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 e documentação da fase de entendimento do negócio.

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

- **Entradas:** Dados brutos da base de dados escolhida, critérios de seleção de variáveis.
- **Saídas:** Conjunto de dados preparados e limpos para a modelagem.
- **Dependências:** Essa fase depende da conclusão da fase de entendimento dos dados para realizar a melhor preparação possível para a fase seguinte de modelagem.

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 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 modelagem.

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. Além

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

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 classificadores 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

A coleta inicial de dados é uma fase crucial para o sucesso do projeto de mineração de dados. A aquisição dos dados listados nos recursos do projeto visa garantir a disponibilidade das informações necessárias para alcançar os objetivos comerciais.

2.1. Coletar dados iniciais

2.1.1. Relatório inicial de produção de dados

O relatório inicial de produção de dados descreve minuciosamente os dados utilizados, seus requisitos de seleção e a avaliação da importância relativa dos atributos. Essa análise se estende à qualidade dos dados nas fontes individuais e nos dados resultantes da fusão de diferentes fontes, identificando possíveis problemas, como inconsistências.

Esta etapa enfatiza a necessidade de compreender profundamente os requisitos de dados, selecionar atributos relevantes e aplicar estratégias eficazes para lidar com desafios específicos dos dados.

Planejamento de Requisitos de Dados:

Ao planejar os requisitos de dados, é essencial identificar de maneira precisa as informações necessárias para atender aos objetivos de mineração de dados. Isso envolve uma compreensão clara das questões de negócios traduzidas em termos técnicos. Além disso, é crucial realizar uma verificação rigorosa para garantir a real disponibilidade de todas as informações necessárias. Isso inclui a confirmação de que os dados requeridos estão acessíveis e podem ser integrados de maneira eficiente no projeto.

Critério de Seleção:

A definição de critérios de seleção é uma etapa estratégica que determina quais atributos são essenciais para os objetivos específicos de mineração de dados. Isso requer uma análise cuidadosa dos dados disponíveis, identificando quais variáveis são relevantes para a construção do modelo preditivo. A seleção de tabelas e arquivos apropriados baseia-se nos critérios estabelecidos, garantindo que apenas os dados relevantes sejam incluídos. Além disso, é importante determinar a quantidade necessária de dados, considerando o período histórico relevante para o exercício. Essa fase envolve decisões estratégicas sobre a abrangência temporal dos dados, alinhadas aos objetivos do projeto.

Inserção de Dados:

A inserção de dados aborda questões práticas relacionadas à codificação de entradas de texto livre, quando aplicável. Isso envolve a transformação de dados não estruturados em formatos que possam ser facilmente incorporados nos modelos de mineração. Além disso, a estratégia para lidar com atributos ausentes é fundamental para manter a integridade dos dados.

Essa etapa pode envolver a aplicação de técnicas avançadas de preenchimento ou imputação de dados para garantir a completude e relevância do conjunto de dados. Explorar técnicas avançadas de extração de dados também é uma consideração importante para melhorar a qualidade e a abrangência dos dados coletados. Isso pode incluir a análise de fontes não eletrônicas e a incorporação de conhecimentos externos para enriquecer os dados.

Referências Bibliográficas

Valentim Realinho, Jorge Machado, Luís Baptista, & Mónica V. Martins. (2021). Predict students' dropout and academic success (1.0) [Conjunto de dados]. Zenodo. https://doi.org/10.5281/zenodo.5777340