CENTRO UNIVERSITÁRIO UNIFACVEST

CIÊNCIA DA COMPUTAÇÃO

GLÉDISON DE AVILA BOMFIM

EVASÃO ESCOLAR EM INSTITUIÇÃO PRIVADA DE ENSINO SUPERIOR: ANÁLISE E PREDIÇÃO

LAGES
2021

GLÉDISON DE AVILA BOMFIM

EVASÃO EM UMA INSTITUIÇÃO PRIVADA DE ENSINO SUPERIOR: ANÁLISE E PREDIÇÃO

Trabalho de Conclusão de Curso apresentado ao Centro Universitário Unifacvest como requisito básico para a aprovação no curso de Ciência da Computação (TCC II).

Orientador (a):

Professora ME. Juliana Facchini de Souza

Co-orientador:

Professor ME. Marcio José Sembay

LAGES
2021

GLÉDISON DE AVILA BOMFIM

EVASÃO EM UMA INSTITUIÇÃO PRIVADA DE ENSINO SUPERIOR: ANÁLISE E PREDIÇÃO

Trabalho de Conclusão de Curso apresentado ao Centro Universitário Unifacvest como requisito básico para a aprovação no curso de Ciência da Computação (TCC II).

Orientador (a):

Professora ME. Juliana Facchini de Souza

Co-orientador:

Professor ME. Marcio José Sembay

Lages, SC 12/07/2021 Nota__ _____

_____

Marcio José Sembay

LAGES
2021

AGRADECIMENTOS

Agradeço primeiramente aos meus pais, que sendo meus pais adotivos me ensinaram o verdadeiro significado de amor e que sempre estiveram ao meu lado, me dando segurança e apoio.

A minha família, em especial meu falecido avô João, que nos deixou muitos ensinamentos, foi um grande homem.

Aos amigos de Lages, que estiveram ao meu lado, foram minha família quando eu estava distante da minha.

A minha professora, orientadora e amiga Juliana Facchini, que desde o início do curso de graduação tive muita admiração e carinho, obrigado por confiar em mim, me apoiar e chamar a minha atenção quando necessário.

Ao Centro Universitário Unifacvest, instituição que tive a oportunidade de trabalhar.

Aos professores, que não mediram esforços para nos transmitir conhecimento e nos auxiliar, vocês foram fundamentais para a conclusão deste projeto.

Aos meus colegas, em especial aqueles que estiveram comigo neste período de graduação.

Dedico este trabalho aos meus pais Jurema e Alberi que, com muito esforço e determinação, fizeram com que eu tivesse a oportunidade de ser quem eu sou hoje, sem eles eu jamais teria chegado aqui.

RESUMO

O alto índice de desemprego, aliado com a grande falta de mão obra qualificada no mercado de trabalho fez com que despertasse o interesse em buscar informações significativas para entender os motivos que levam os alunos evadirem de seus cursos de graduação, mesmo sabendo que no mercado de trabalho quem possui ensino superior completo consegue alcançar melhores posições e salários. Ao mesmo tempo, essa evasão causa um prejuízo de cerca de 600 milhões de reais ao ano no Brasil às instituições de ensino.

A análise de dados é uma das áreas que mais crescem atualmente. As empresas entendem que há grande valor nas informações extraídas de suas bases de dados, que muitas vezes trazem respostas que auxiliam na tomada de decisão e possibilita a empresa ser mais assertiva.

Este trabalho propõe realizar o processo de extração, transformação e carregamento de dados de uma instituição de ensino, gerando datasets, realizando análises exploratórias e entrega de predições baseadas no perfil de evasões da instituição superior privada. Possibilitando assim, que os administradores busquem insights e desenvolvam medidas preventivas à evasão.

Palavras-chave: Análise. Predição. Evasão.

ABSTRACT

The high rate of unemployment, combined with the great lack of qualified labor in the labor market, has aroused interest in seeking significant information to understand the reasons that lead students to drop out of their undergraduate courses, even knowing that in the market of Whoever has completed higher education is able to achieve better positions and salaries. At the same time, this evasion causes losses of around 600 million reais a year in Brazil to educational institutions.

Data analysis is one of the fastest growing areas today. Companies understand that there is great value in the information extracted from their databases, which often provide answers that help in decision making and enable the company to be more assertive.

This work proposes to carry out the process of extracting, transforming and loading data from an educational institution, generating data sets, performing exploratory analyzes and delivering predictions based on the profile of evasion of the private higher institution. Thus enabling administrators to seek insights and develop measures to prevent evasion.

Keywords: Analysis. Prediction. Evasion.

LISTA DE FIGURAS

LISTA DE GRÁFICOS

LISTA DE TABELAS

LISTA DE SIGLAS

INEP	Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira
MEC	Ministério da Educação
SEMESP	Sindicato das Mantenedoras de Ensino Superior
IES	Instituição de Ensino Superior
CRM	Customer Relationship Management
ERP	Enterprise Resource Planning
IoT	Internet das coisas
CRISP-DM	Cross Industry Standard Process for Data Mining
DDD	Data Driven Decision
MIT	Instituto de Tecnologia de Massachusetts
XML	Extensible Markup Language
JSON	JavaScript Object Notation
NoSQL	Not Only SQL
HTTP	HyperText Transfer Protocol
HTML	HyperText Markup Language

SUMÁRIO

1. INTRODUÇÃO

O mercado de trabalho torna-se cada vez mais competitivo e as empresas consequentemente, mais exigentes com o nível de qualificação de seus colaboradores. Nessa realidade, o diploma de graduação passou a ser um grande diferencial para alcançar melhores colocações. Pesquisa realizada em 2018 pelo IBGE(Instituto Brasileiro de Geografia e Estatística), apontou que o salário médio mensal de quem possui ensino superior completo é de R$5.969,32, já daqueles que não possuem a média é de R$2.020,88. Atualmente, apenas 17,4% da população brasileira possui o ensino superior completo.

Apenas com o intuito de comparação, uma pesquisa realizada nos Estados Unidos, pelo Bureau of Labor Statistics, apontou que os cidadãos dos Estados Unidos com diploma de bacharel embolsam cerca de US $1.173 em média a cada semana. Já aqueles com apenas diplomas do ensino médio ganham uma média de apenas US $712 por semana.. Ou seja, em ambos países nota-se a importância do ensino superior completo para a obtenção de uma renda maior.

Mesmo a par dessa realidade, nota-se um alto índice de abandono da graduação pelos alunos.No Brasil, dados do INEP(Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira) indicam que, em 2016, a taxa de concluintes em instituições de ensino superior privadas foi de aproximadamente 14,3% dos ingressos, sendo que na região sul o índice atingiu 14,1%. Já em 2017, apenas 14,5% dos estudantes concluíram o curso superior, sendo que na região sul do país, a taxa permaneceu em torno de 14,5%.

Smith e Naylor (2011) apresentam que, enquanto aproximadamente 37% dos estudantes evadiram do ensino superior nos Estados Unidos, na Inglaterra essa taxa ficou próxima dos 18%. Ou seja, a taxa de evasão no Brasil é bastante alta em comparação aos países citados.

O gráfico 1, mostra que o número de matrículas cresceu no período analisado, porém, o número de concluintes se manteve abaixo de um milhão no mesmo período.

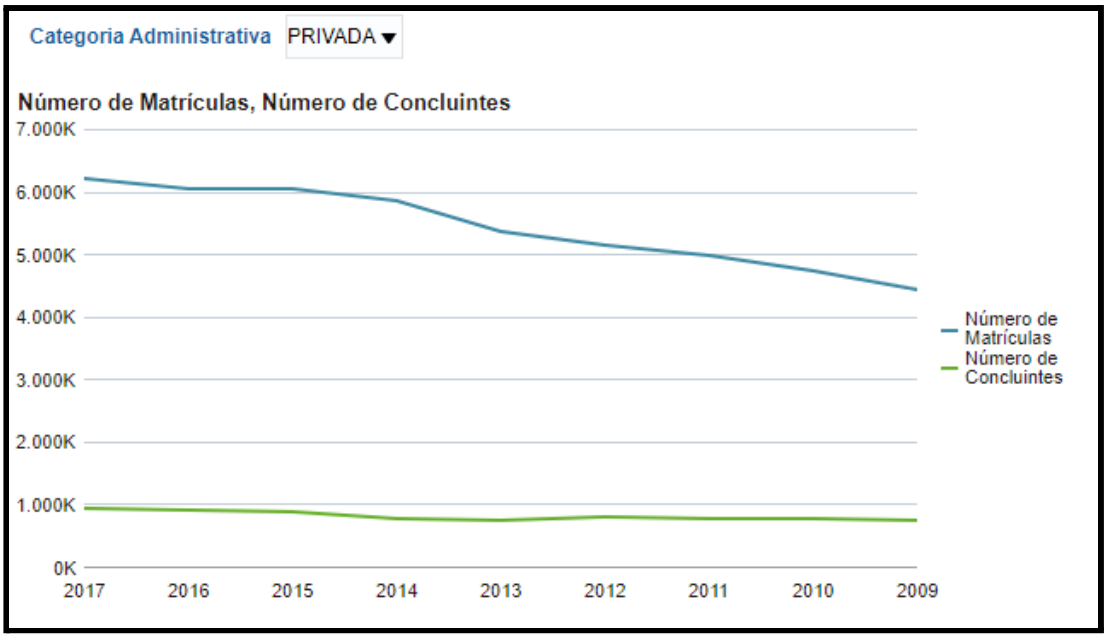


Gráfico 1- Número de matrículas por número de concluintes
fonte: https://inepdata.inep.gov.br/

Atualmente, o país conta com 2639 instituições de ensino superior privadas, sendo que 108 delas se encontram no estado de Santa Catarina.(MEC, 2021)

A média da mensalidade de um curso presencial é de R $1,2 mil, somando cerca de R$ 14,4 mil ao ano, já nos Estados Unidos o valor médio anual é cerca de U$ 35,8 mil, se converter o valor médio de mensalidades do Brasil para a moeda estrangeira, levando em consideração o valor de câmbio U$ 1 é equivalente à R$ 5, teríamos um valor médio ao ano de U$ 2,8 mil. Ou seja, o custo médio anual brasileiro é equivalente a uma mensalidade média de graduação nos Estados Unidos. Esses dados colocam em dúvida, então, o quanto o valor do curso influencia a alta evasão de alunos.

Como podemos ver na tabela 1, a relação entre os números de matrículas e o número de concluintes é muito distante.

Ano Censo	Nome Região	Número de Vagas Oferecidas	Número de Tipo de Ingressos	Número de Matrículas	Número de Concluintes	Número de Cursos	Número de Inscritos
2017	BRASIL	6.075.252	2.135.126	6.529.681	947.606	33.522	14.605.840
	CENTRO-OESTE	554.885	208.370	607.357	89.956	3.072	1.306.750
	NORDESTE	1.138.928	441.493	1.447.248	188.730	6.715	3.804.497
	NORTE	383.771	143.622	473.716	64.851	2.513	1.367.659
	SUDESTE	3.086.636	1.026.159	3.003.075	458.824	14.861	6.516.557
	SUL	911.032	315.482	998.285	145.245	6.361	1.610.377
2016	BRASIL	6.180.251	2.632.500	6.554.283	938.732	32.959	14.407.344
	CENTRO-OESTE	554.091	264.225	606.523	85.890	2.969	1.279.507
	NORDESTE	1.134.737	573.360	1.444.368	179.953	6.417	3.827.094
	NORTE	416.474	203.649	473.479	70.367	2.475	1.397.271
	SUDESTE	3.202.510	1.181.240	3.020.865	460.629	14.849	6.166.490
	SUL	872.439	410.026	1.009.048	141.893	6.249	1.736.982

Tabela 1: Dados coletados no censo 2017 fonte: https://inepdata.inep.gov.br/

2. JUSTIFICATIVA

Entre 2014 e 2017 aproximadamente 2,56 milhões de pessoas ingressaram na rede privada brasileira, sendo que 56,8% destas evadiram. (SEMESP, 2019)

Se levarmos em consideração o valor médio de uma mensalidade no curso presencial, as instituições privadas, deixaram de faturar aproximadamente R$17,5 milhões de reais neste período.

Portanto, é de interesse dessas instituições buscar formas de manter seus alunos devidamente matriculados e ativos em seus cursos. Assim, há um crescimento no interesse de ferramentas que permitem a melhoria de processos administrativos, educacionais e também de comunicação com os discentes.

A análise de dados mostra-se de extrema importância para que os administradores tenham em mãos uma ferramenta que os auxiliem a entender o perfil de seus clientes, principalmente aqueles com maior chance de evasão, baseados nos dados que são gerados através dos ambientes digitais e a partir daí, buscar melhorias que diminuam essa taxa.

3. OBJETIVOS

3.1. Geral

O referente Trabalho de Conclusão de Curso possui como principal objetivo identificar e apresentar indicadores que influenciam na evasão acadêmica no ensino superior privado, bem como realizar análise preditiva, para que a instituição desenvolva ações para aumentar a retenção de alunos.

O processo de análise desenvolvido, realiza o tratamento dos dados do sistema, tornando-os confiáveis e informativos, com o intuito de utilizá-los para a definição de ações e métricas propostas pelos administradores das instituições.

3.2. Específicos

(a) Realizar processo de mineração de dados;
(b) Realizar a análise exploratória;
(c) Responder perguntas baseadas na regra de negócio;
(d) Realizar a predição do perfil de evasão.

4. REFERENCIAL TEÓRICO

4.1 Evasão escolar

É considerado evasão escolar a saída antecipada, antes da conclusão do ano, série ou ciclo, por desistência (independentemente do motivo). Ou seja, quando os alunos iniciam seus cursos mas em algum momento antes de se tornar concluinte, os cessa.

No âmbito da gestão escolar, a evasão indica o grau de sucesso do sistema de ensino. Deve ser levado em consideração os motivos que resultaram na evasão acadêmica do discente, muitas vezes a desistência implica a trajetória acadêmica fragilizada.

Segundo Baggi e Lopes(2011, p. 356), "É um problema que vem preocupando as instituições de ensino em geral, sejam elas públicas ou particulares, pois a saída de discentes provoca graves consequências sociais, acadêmicas e econômicas."

Há diversos fatores que agregam para que o discente se torne uma evasão, estes fatores divergem de acordo com a geografia, quadro econômico e social do ambiente que está inserido.

4.2 Estratégias para reter a evasão

O aumento do número de matrículas e a retenção destes alunos na instituição até o final de seu ciclo educacional são os principais objetivos das instituições de ensino superior, com isso se torna um grande desafio elaborar estratégias e métricas para que isso aconteça. Alguns fatores corroboram para a evasão, como a frustração pelo curso escolhido, a má administração por parte da instituição, estrutura curricular frágil e também, fatores pessoais.

O custo para captação de novos alunos pode chegar a 4 vezes o valor para manter um ativo, sendo então de extrema importância que as IES(instituição de ensino superior) ofereçam condições necessárias para a permanência dos acadêmicos.

Acompanhamento acadêmico e pedagógico adequado, poderiam melhorar os resultados de alunos com baixo rendimento escolar. Para Hipólito(2012) esse seria um dos motivos que fazem que países como Japão, Finlândia e Suécia tenham baixíssimas taxas de evasão.

Algumas estratégias que as IES podem utilizar para reter seus alunos incluem estruturar a gestão acadêmica baseada em indicadores, elaborar e gerir condições especiais aos discentes, investir em tecnologias que contribuam para melhoria de processos acadêmicos, entre outros.

Além das possíveis estratégias citadas anteriormente, há outras ações institucionais que aumentam a retenção escolar, como a qualidade acadêmica, desenvolvimentos de projetos práticos, interativos, atrativos e motivadores e a utilização de informações coletadas através dos dados gerados eletronicamente pelos discentes.

4.3 Crescimento de dados

A difusão dos dispositivos eletrônicos e sua imensa e constante coleta de informações, enviando para servidores de empresas e governos, permitiu o acesso a um grande volume de dados, e consequentemente, um avanço no desenvolvimento computacional que permitisse o armazenamento desses.

Atualmente, a informação é considerada um dos bens mais valiosos do mercado, não apenas para o crescimento de uma nação, como também, a principal estratégia de negócio de muitas corporações. Para que uma empresa seja bem sucedida é imprescindível que ela saiba utilizar as informações extraídas dos dados coletados diariamente.

> De acordo com Sembay(2009) "O avanço das tecnologias de transmissão de informação, novas ferramentas propiciaram novas possibilidades de uso, disseminação e acessibilidade às informações, que para as organizações e instituições de ensino são subsídios imprescindíveis para a inteligência competitiva e os processos decisórios."

Atualmente, há muita facilidade em adquirir dados, podem ser dados internos, como dados de sistemas de gerenciamento da empresa, CRM(Customer Relationship Management) ou, ERP(Enterprise Resource Planning), arquivos gerados pelos colaboradores, sensores físicos, *logs* de sistema e de servidores, como também de fontes externas, através de sistemas públicos, pesquisas e empresas privadas.

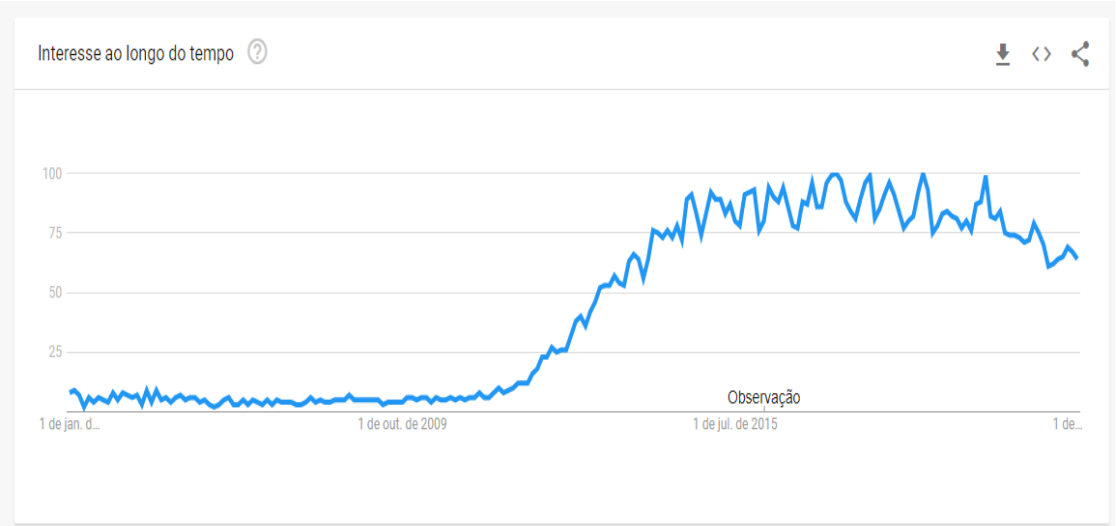


Gráfico 2 - Buscas pelo termo *Big Data* - fonte:
https://trends.google.com/trends/explore?date=all&q=big%20data

4.3 ***Big Data***

O termo inglês, *Big Data*, é utilizado principalmente para descrever um grande volume de dados digitais, principalmente dados não estruturados. Diferente dos tradicionais bancos de dados, estes dados não estruturados permitem que seja mais fácil armazenar publicações de redes sociais como o Facebook, Twitter, bem como, vídeos do YouTube, geolocalização e dados comportamentais gerados digitalmente.

Importante ressaltar que *Big Data*, não é só um grande volume de dados, há mais duas propriedades que devem ser consideradas, a variedade e a veracidade dos dados.

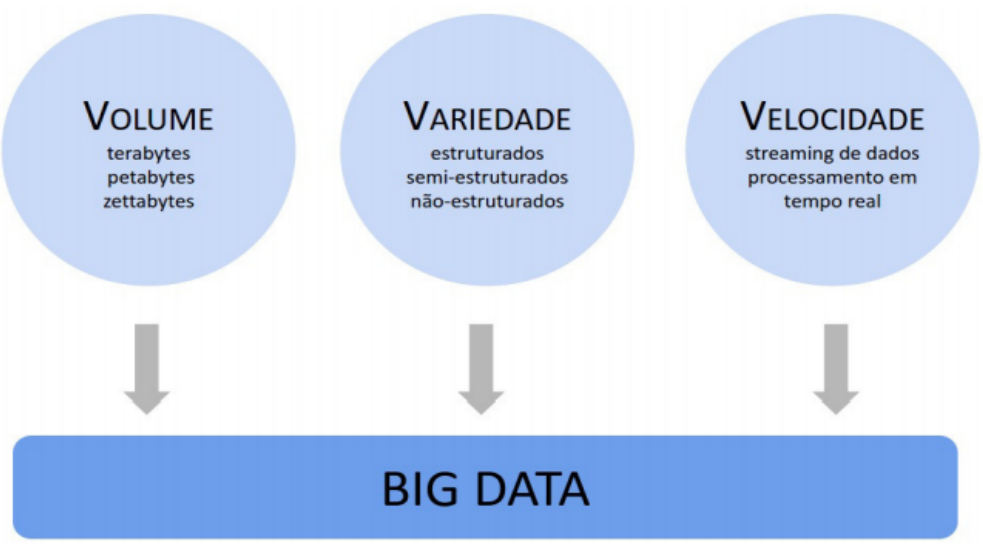


Figura 1 - Buscas pelo termo *Big Data* -
fonte: https://docplayer.com.br/113530610-Introducao-a-big-data-juciander-l-moreira-wallace-brito.html

> “Big data é um termo aplicado a conjuntos de dados cujo tamanho ou tipo está além da capacidade de bancos de dados relacionais tradicionais de capturar, gerenciar e processar os dados com baixa latência.”(IBM, 2021)

Os dados são considerados o novo petróleo e com isso permite elevado grau de capacidade estratégica e analítica para tomada de decisão baseada em dados e, desta forma, gerando valor para empresas e instituições.

4.5 Coleta de dados

Há diversos métodos para coletar dados. Estes, consistem em captar dados das mais diversas fontes, como sistemas, aplicativos, formulários, comportamento do usuário na internet, assim como através de dispositivos conhecidos como *IoT*(Internet das coisas).

Empresas entendem que através dos dados é possível extrair informações que podem se tornar estratégias corporativas para melhorar a tomada de decisões. Os dados a serem utilizados precisam ser confiáveis, pois ao contrário, as informações extraídas seriam apenas um monte de informações sem valor real.

Para que não haja problemas como o citado anteriormente, é necessário que os engenheiros de dados façam a coleta e preparação de dados de forma que permita a confiabilidade neles.

4.6 Mineração de dados

O conceito principal de mineração de dados é a extração de conhecimento útil a partir dos dados coletados para resolver problemas de negócios, com processos sistemáticos unidos à ferramentas computacionais, evoluindo o processo entre etapas previamente definidas.(FAYYAD, Usama, 1996)

Das inúmeras metodologias de mineração de dados, destacamos a Cross-Industry Standard Process of Data Mining (CRISP-DM). A proposta desta metodologia é a definição de um processo aplicável a todos os possíveis segmentos da indústria, tornando-se mais ágil, com menor custo e facilidade de gerenciamento, no processo de mineração de dados.

A Figura 2 ilustra a metodologia CRISP-DM em seu formato cíclico e de forma implícita. Cada fase é composta por tarefas que precisam ser realizadas para que seja possível avançar para a próxima. Por mais que esta metodologia possua um padrão de fases no processo de modelagem de dados, a mesma não impede que uma seja adicionada no meio do fluxo, assim como novas variáveis ou até mesmo novos

dados, desde que estejam aptos a serem adicionados.

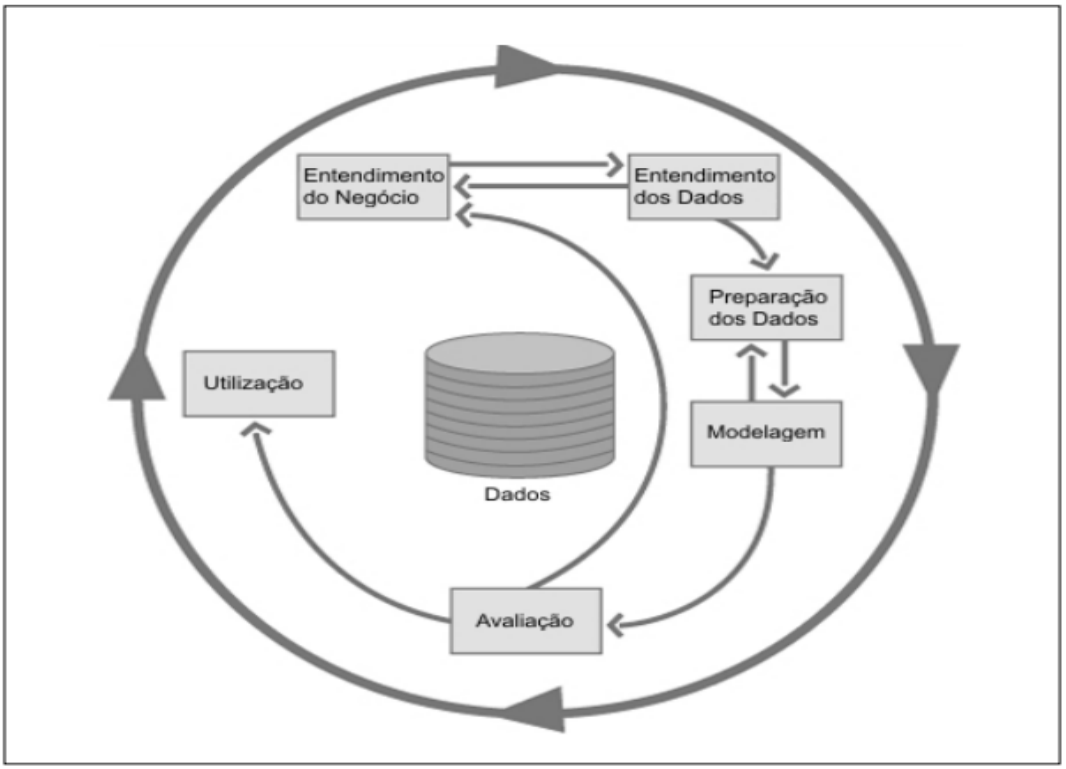


Figura 2: Metodologia CRISP-DM para mineração de dados
fonte:<https://www.datascience-pm.com/crisp-dm-2/>

A metodologia CRISP-DM é composta por 6 fases, sendo elas:

- **Entendimento do Negócio**: tem por objetivo assimilar os objetivos do projeto e dos requisitos, de acordo com o ponto de vista do negócio. Aqui, o problema de mineração de dados é definido e também projeta-se as metas pretendidas;

- **Entendimento dos Dados**: inicia-se com um conjunto de dados e procede com atividades que buscam familiarizar-se com eles, identificar problemas de qualidade, identificar discernimento ou descobrir sub-conjuntos interessantes que possam formar hipóteses sobre a informação oculta;

- **Preparação dos Dados**: inclui todas as atividades de construção do conjunto de dados(dataset) a ser utilizado. As tarefas de preparação de dados são, quase sempre, executadas inúmeras vezes até que se obtenha o resultado necessário para que se possa avançar para a etapa seguinte. As tarefas incluem a seleção de tabelas, registros e atributos, bem como a transformação dos dados e limpeza, para que seja

possível executar algoritmos de mineração;

- **Modelagem**: um conjunto de técnicas de modelagem são selecionadas e aplicadas, com ajustes de parâmetros para valores ótimos. Geralmente há várias técnicas para o mesmo tipo de problema. Algumas delas possuem requisitos específicos para que seja realizada a formação de dados, desta forma, retornar à fase de preparação é frequente no processo;

- **Avaliação**: o modelo construído na fase anterior é avaliado e são revisados os passos executados em sua construção, visando atender os requisitos definidos na primeira fase. O objetivo principal é determinar se existe alguma questão de negócio que não foi suficientemente considerada;

- **Implementação do Modelo**: posterior a construção e avaliação do modelo, a fase de utilização pode ser a geração de dados apresentáveis na forma de relatórios de suporte à tomada de decisão, ou até mesmo, submeter o modelo a processo de ajustes e evolução.

4.7 Decisão orientada a dados

Decisão orientada a dados, também conhecida como DOD ou em inglês DDD (*data driven decision*), é um método de tomar decisões de negócios, não apenas na intuição e experiência dos gestores, mas também na análise de dados. Este método não descarta a necessidade de gestores com grande conhecimento de negócios e sim, se torna um grande poder de decisão quando aliados.

Os benefícios da tomada de decisão orientada a dados têm sido demonstrados conclusivamente. O economista Erik Brynjolfsson e seus colegas do MIT(Instituto de Tecnologia de Massachusetts) e da Penn 's Wharton School realizaram um estudo de como a tomada de decisões baseada em dados afeta o desempenho das empresas. Eles desenvolveram uma medida de tomada de decisões baseadas em dados que classifica as empresas quanto ao uso delas. Eles mostram que, estatisticamente, quanto mais orientada por dados, mais produtiva uma empresa é.(Brynjolfsson, Hitt e Kim, 2011)

No mundo dos negócios, cada problema que seja necessário a tomada de decisão orientada a dados é exclusivo, tendo seu conjunto de objetivos, metas e limitações.

Da mesma forma que acontece em grande parte na engenharia, há um bloco de de tarefas comuns que habita os problemas de negócios. (PROVOST, F.; Fawcett, T. 2016)

4.8 Armazenamento de dados

O armazenamento de dados possui um grande e importante papel no mundo digital. Nenhum recurso computacional e serviços digitais poderiam ser disponibilizados sem a utilização de bancos de dados para receber, armazenar e disponibilizá-los de forma eficaz e segura. Com a crescente disponibilidade de um grande volume de dados, empresas de diversos segmentos passaram a perceber o grande potencial que os mais diversos tipos de dados podem oferecer. Essa realidade, alavancou o desenvolvimento de ferramentas que permitissem o armazenamento e gerenciamento desse grande volume.. Atualmente, oss principais modelos de banco de dados são o Relacional e o Não-Relacional.

O Modelo relacional é o mais utilizado, estando inserido no mercado há mais de 40 anos. Seu modelo de armazenamento é em formato de tabelas que podem estar relacionadas com outras tabelas do mesmo banco , com uma estrutura previamente definida. Antes que possa armazenar qualquer informação em um banco de dados relacional, é necessário que seja definida sua estrutura, sequência, tamanho e o tipo de dados. Uma importante característica do banco relacional é o suporte à propriedade ACID(Atomicidade, Consistência, Isolamento, Durabilidade), que garante a integridade e confiabilidade dos dados através dos seguintes recursos:

- **Atomicidade**: garante que todas as alterações realizadas por uma transação sejam efetivadas no banco de dados, ou nenhuma, se houver alguma situação inesperada. Ou seja, a transação é ou não realizada, não há processos parciais.**Consistência**: prevê que só é possível realizar novas transações que não venham ferir com as regras do banco , permitindo que esteja sempre em estado consistente. **Isolamento**: esta propriedade permite que os eventos realizados em uma transação não interfiram nos eventos de um transação concorrente, desta forma, apenas uma transação por vez irá realizar alterações no dados. **Durabilidade:** garante que o resultado de toda transação executada com sucesso no banco seja mantido de forma segura, mesmo que ocorram falhas.

Mesmo que seja muito eficiente e utilizado em diversos cenários, o banco de dados relacional é projetado para armazenar majoritariamente dados estruturados, ou seja, dados com esquemas bem definidos e adequados para o formato de tabelas. Isto torna-o uma limitação para *Big Data,* já que esse inclui também, dados semi estruturados e não estruturados.

Podemos entender como dados semi estruturados, aqueles que possuem uma estrutura pré-definida, mas não com o mesmo rigor dos dados relacionais. Estas estruturas são muito utilizadas como meio de marcação dos dados, sendo os formatos mais utilizados JSON(JavaScript Object Notation) e XML(eXtensible Markup Language). Os dados não estruturados inclui vídeos, imagens e alguns formatos de textos. Estes tipos de dados geram grande dificuldade de armazenamento e gerenciamento através de modelos relacionais.

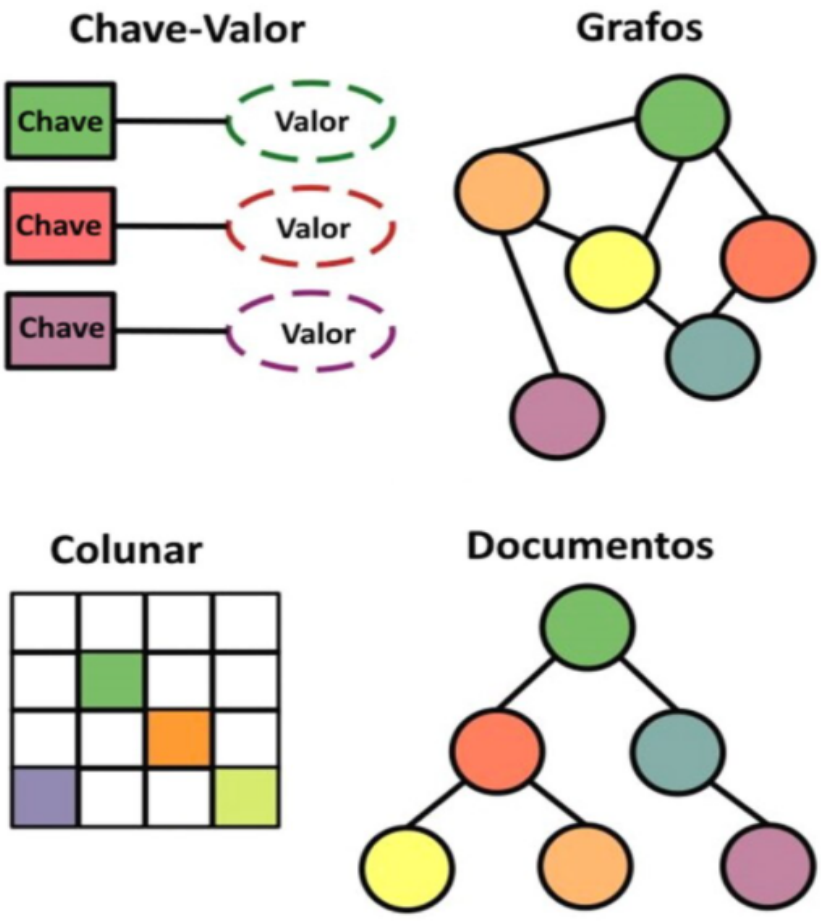


Figura 3: modelos de bancos de dados não relacionais
fonte:<https://blogs.oracle.com/lad-cloud-experts/pt/introducao-ao-oracle-nosql-database-cloud-parte-1>

O m**odelo não relacional** surgiu através da necessidade de gerenciar dados semi estruturados e não estruturados. Seu maior objetivo é oferecer flexibilidade quanto a sua estrutura.

Esses modelos foram criados para atender às necessidades de flexibilidade, disponibilidade, escalabilidade e desempenho provenientes do *Big Data.* Ao contrário do banco de dados relacional em que o principal objetivo é a integridade dos dados, os modelos não relacionais, tendem a sacrificar uma ou mais propriedades ACID, para que assim possa oferecer maior desempenho e escalabilidade para as soluções que trabalham com grande volume de dados. De tal forma que não existe um padrão único que atenda todos os possíveis cenários. No mundo dos modelos não relacionais, os principais são: o modelo orientado a chave-valor, orientado a documentos, orientado a colunas e orientado a grafos.

4.8.1 O banco de dados orientado a chave-valor é o banco de dados NoSQL que possui a estrutura mais simples. Ele possui como estratégia de armazenamento de dados a utilização de chaves como identificadores das informações gravadas em um campo de valor. A chave normalmente é composta por um campo do tipo *String*. Já o campo valor pode ter diferentes tipos de dados, sem precisar ter um esquema previamente definido. Este modelo pode ser tanto usado para gravar seus dados em um banco de dados, quanto para utilização em memória, permitindo um acesso mais ágil às informações.

4.8.2 O banco de dados orientado a documentos é considerado uma extensão do orientado a chave-valor, sendo este o mais popular. Ele oferece simplicidade e flexibilidade no gerenciamento dos dados. Também oferece meios para a criação de índices sobre os valores dos dados armazenados, enriquecendo as possibilidades de consultas. Podemos definir os documentos como estruturas flexíveis que podem ser geridos através de dados semi estruturados, como o *XML* e *JSON*.

4.8.3 O modelo orientado a colunas trata-se do mais complexo de todos. É considerado uma extensão do armazenamento orientado a chave-valor e possui conceitos similares ao do modelo relacional, como a criação de linhas e colunas. Mas há diferenças fundamentais na forma em que as estruturas são criadas. No banco de dados orientado a colunas, o responsável pela modelagem define o que é chamado de "famílias de colunas", elas são organizadas em grupos de itens de dados que são frequentemente utilizados em conjunto. No registro de um item pode ter informações

gravadas em diversas famílias de colunas, que podem estar armazenadas em servidores distintos.Esta forma de armazenamento fornece flexibilidade e escalabilidade.

4.8.4 O banco de dados orientado a grafos é útil quando a descoberta de como os dados estão relacionados é mais importante do que os dados em si. Entre os quatro tipos de banco não relacionais citados, o orientado a grafos é o mais especializado. Este possui uma estrutura definida na teoria dos grafos, utilizando vértices e arestas para armazenar os dados dos itens coletados e os relacionamentos entre eles. Este modelo é muito utilizado em aplicações que traçam os caminhos existentes nos relacionamentos entre os dados.

4.9 Análise de dados

Análise de dados é a habilidade de utilizar dados, extrair informações baseadas em análises e utilizar processos sistemáticos para conduzir a uma tomada de decisão mais eficiente.(MACHADO, 2018)

Quando se tem milhares de informações à sua disposição, é fundamental saber o que é necessário extrair de um montante de dados, que às vezes, não possuem conexões entre si de forma direta ou explícita. Ou seja, unir poder computacional com perguntas determinadas pela regra de negócio da organização detentora das informações.

É fundamental que os analistas sigam processos bem definidos para que a análise tenha confiabilidade e possa ser usada para tomada de decisões. Reconhecer a base de dados, conhecer os atributos, variáveis e suas condições são tarefas obrigatórias na análise.

Há diferentes tipos de análise de dados, alguns deles são:

4.9.1 Análise descritiva: é baseada em fatos. Este tipo de avaliação é realizada através de resultados obtidos, sendo a técnica mais utilizada pelas empresas, seu principal objetivo é responder a pergunta: “O que aconteceu?”. Os indicadores são levantados a partir dos dados gerados pela empresa. Normalmente é utilizado *dashboards* para apresentar as informações obtidas.

4.9.2 Análise diagnóstica: tem como objetivo identificar relações de causa e efeito para explicar um acontecimento, seu foco é responder a pergunta: “O que aconteceu?”. Esta análise possui relação direta com a análise descritiva, necessária para avaliar projetos já executados.

4.9.3 Análise preditiva: o mais popular tipo de análise, sua essência é a previsão de cenários futuros baseados em padrões revelados através dos dados. A análise preditiva não possibilita prever o que vai acontecer, mas sim, o que vai acontecer se determinadas condições se cumprirem. Ela permite não apenas compreender o passado, mas também oferece a possibilidade de obter informações sobre "o que pode acontecer" no futuro, tanto em relação aos riscos como
também oportunidades. É utilizado mecanismos de aprendizagem de máquina e técnicas estatísticas, para identificar padrões, tendências e exceções nos dados históricos, e com isso, criar modelos que permitam realizar predições.

4.9.4 Análise prescritiva: tendo como base a análise preditiva, é o processo em que se recomenda algo potencialmente previsto, visando direcionar para decisões assertivas. Os algoritmos de análise prescritiva são programados com baixa intervenção humana, para que o algoritmo seja capaz de se adaptar de acordo com os parâmetros recebidos por ele, buscando uma otimização automática. Para que seja possível realizar essas decisões automaticamente é necessário um grande volume de dados. (CASTRO e FERRARI, 2016)

4.10 Visualização de dados

Trata-se de representações visuais de dados, apresentados em sua maioria por gráficos. De acordo com Knaflic(2018), os gráficos para apresentação de análises realizadas em dados, devem ser estudados e elaborados de acordo com o seu público alvo, levando em consideração a técnica do “Quem, o quê e como?”

O processo de visualização de dados é uma etapa extremamente importante, pois, aliada com a regra de negócio, responde às perguntas levantadas no processo de análise, de forma que os insumos gerados estejam alinhados com a realidade da organização solicitante.

4.11 Aprendizado de máquina

Também conhecido pelo termo em inglês *Machine Learning,* é a programação de computadores para que eles possam aprender, identificar padrões e tomar decisões com os dados. Os modelos de *Machine Learning* se destacam muito com problemas complexos para abordagens tradicionais, ou, que não possuam algoritmos prontos para a solução.

Os modelos de Machine Learning, são classificados de acordo com a quantidade

e tipo de supervisão que recebem durante o seu treinamento. Existem quatro categorias principais, supervisionado, não supervisionado, semi supervisionado e por reforço.

No Aprendizado supervisionado, os dados de treinamento fornecidos ao modelo incluem as soluções desejadas, conhecido por rótulos, muito utilizados para tarefas de classificação.

Já no aprendizado não supervisionado, os dados de treinamento não possuem rótulos, ou seja, o modelo irá tentar aprender de forma independente. Este modelo é bastante utilizado para identificação de padrões em dados desconhecidos.

O modelo semi supervisionado, consegue trabalhar com uma pequena parte de seus dados rotulados, e uma grande quantidade não rotulada, muito utilizado para detecção de pessoas em imagens.

O aprendizado por reforço é o que menos tem semelhança com os demais, este modelo observa, seleciona e executa ações para obter recompensas em troca, ou penalidades em forma de recompensas negativas. Ele deve aprender por si só qual é a melhor estratégia, chamada de política, para obter o maior número de recompensas ao longo do tempo.(GÉRON, A.; 2019)

4.12 Árvores de decisão

São algoritmos de Aprendizado de Máquina Supervisionados que podem executar tarefas de classificação e regressão. Estes algoritmos são muito poderosos e possuem capacidade de moldar conjuntos complexos de dados. Também são os componentes fundamentais das Florestas Aleatórias, que estão entre os algoritmos mais poderosos da atualidade. Uma grande vantagem das Árvores de decisão é que elas exigem pouca preparação de dados.

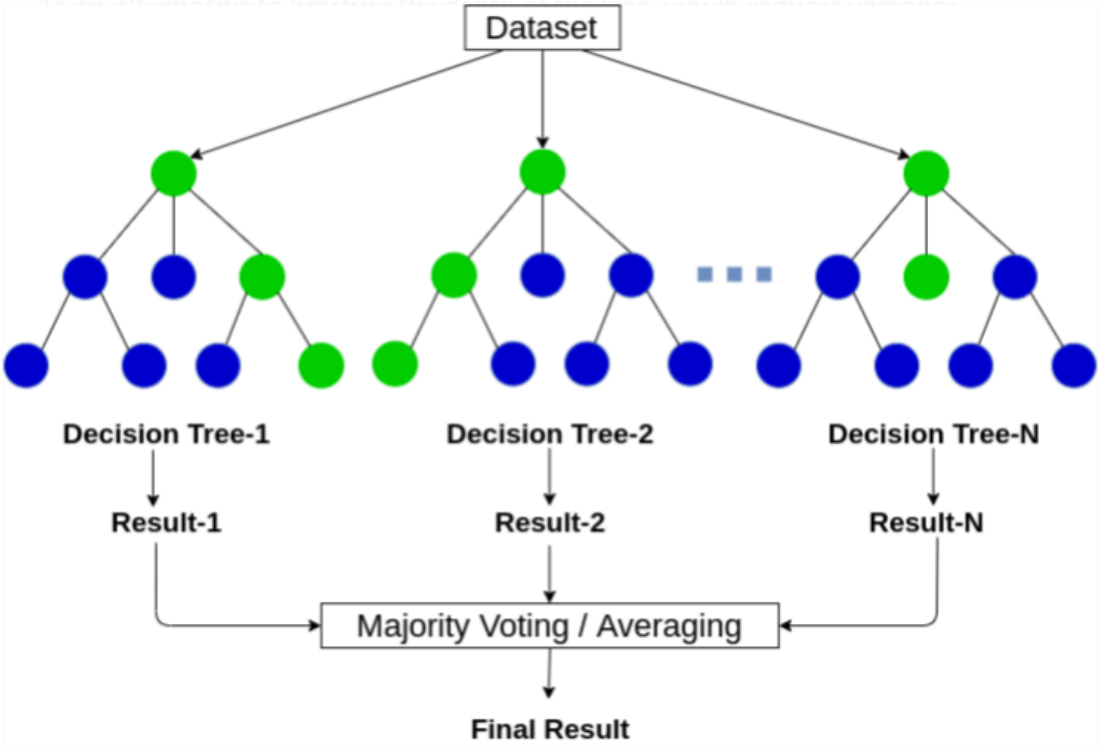


Figura 4: *Random forest*
fonte:<Random Forest (Easily Explained). (With Python implementation in depth!) | by Shubham Gupta | Medium>

4.13 Ensemble Learning

O termo em inglês *Ensemble Learning* é um conjunto de previsões, podendo ser classificadores ou regressores, que geralmente alcançam resultados melhores do que o melhor previsor individual.

Há três métodos Ensemble mais conhecidos, Bagging, Boosting e Stacking.

4.13.1 O *Bagging*, abreviação para *bootstrap aggregating,* utiliza o mesmo algoritmo de treinamento para cada previsor, mas treina-os em diferentes subconjuntos aleatórios, realizando amostragem com substituição. Este método permite que as instâncias de treinamento sejam estudadas diversas vezes pelo mesmo previsor. Uma grande vantagem do método *bagging* é seu alto poder de escalonamento.

4.13.2 O método *Boosting*, originalmente denominado *hypothesis boosting*, refere-se a qualquer método *ensemble* que combina vários aprendizes fracos para formar um forte. O objetivo principal dos métodos *boosting* é treinar sequencialmente os previsores, cada um tentando corrigir seu antecessor. Os métodos *boosting* mais conhecidos são *AdaBoost* e *Gradient*. A principal característica do

AdaBoost é que ele utiliza um novo previsor para corrigir o peso relativo de seus antecessores e com isso resulta em novos previsores, focando em casos mais difíceis. Assim como *AdaBoost*, o *Gradient* adiciona novos previsores sequencialmente a um conjunto, cada um corrigindo o seu antecessor. Entretanto, em vez de ajustar os pesos da instância a cada iteração, o *Gradient* tenta ajustar o novo previsor aos erros residuais feitos pelo previsor anterior.

4.13.3 O *Stacking*, abreviação de *stacked generalization,* é baseado em uma ideia simples: em vez de utilizar funções triviais para agregar as previsões de todos os previsores em um conjunto, após cada previsor entregar seu resultado, um previsor final, denominado *blender,* escolhe as melhores previsões e faz a final, através de uma votação.(GÉRON, A.; 2019)

4.13.4 O algoritmo Random Forest, ou Floresta Aleatória, utiliza uma combinação de múltiplas árvores de decisão. Estes conjuntos de dados, são amostradas aleatoriamente n instâncias, com possibilidade de repetição, sendo n o número de instâncias do conjunto de dados original. Cada árvore é induzida a partir de um desses subconjuntos, porém cada nó da árvore utiliza m atributos da quantidade total f de atributos, sendo $m = 1$ quando $f = 1$, e $m < f$ quando $f > 2$. Os m atributos são escolhidos aleatoriamente e com repetição.

É necessário definir a quantidade de árvores de decisão presentes na Random Forest. Através do uso combinado dessa variedade de árvores de decisão é possível convergir o valor de erro para um valor que não sofreu overfit, ou sobre-ajuste, em relação ao conjunto de dados fornecido. (GÉRON, A.; 2019)

5. APLICAÇÃO

A aplicação desenvolvida neste trabalho teve como seu principal objetivo ser um módulo que pudesse ajudar administradores de instituições de ensino superiores privadas nas tomadas de decisões, baseando-se em dados gerados pelos próprios alunos.

Este projeto foi elaborado com metodologias e ferramentas de pontas, sendo elas, utilizadas pelas maiores empresas do mundo e que possuem suas diretrizes e tomada de decisões baseando-se em dados, conceito de Data Driven.

O projeto foi desenvolvido com com o propósito de ser validado em uma empresa de software com foco em gestão educacional.

5.1 Python

Python é uma linguagem de programação interpretada, interativa e orientada a objetos. O python incorpora módulos, exceções, tipagem dinâmica, tipos de dados dinâmicos de nível muito alto e classes. Ele oferece suporte a vários paradigmas de programação além da programação orientada a objetos, como programação procedural e funcional. Python combina poder notável com sintaxe muito clara. (PYTHON, 2021)

Possui interfaces para muitas chamadas de sistema e bibliotecas, bem como para vários sistemas de janela, e é extensível em C ou C ++. Também pode ser usado como uma linguagem de extensão para aplicativos que precisam de uma interface programável. Finalmente, o Python é portátil: ele roda em muitas variantes do Unix, incluindo Linux e macOS, e no Windows. (PYTHON, 2021)

O Python tem muita popularidade quando se trata de *data science* e análise de dados, devido ao seu grande número de bibliotecas que em conjunto possibilitam um maior desempenho. As bibliotecas mais famosas do Python para *data science* são:

5.2 NumPy

NumPy é o pacote fundamental para computação científica em Python. É uma biblioteca que fornece um objeto de matriz multidimensional, vários objetos derivados (como matrizes e matrizes mascaradas) e uma variedade de rotinas para operações rápidas , incluindo matemática, lógica, manipulação de forma, classificação, seleção, transformadas discretas de Fourier, álgebra linear básica,

operações estatísticas básicas, simulação aleatória, entre outros.(NumPy, 2021)

5.3 **SciPy**

A biblioteca SciPy é um dos pacotes principais que compõem a pilha SciPy. Ele fornece muitas rotinas numéricas eficientes e fáceis de usar, como rotinas para integração numérica, interpolação, otimização, álgebra linear e estatísticas.(SciPy, 2021)

5.4 Pandas

O Pandas permite trabalhar com DataFrame e possui mecanismos para ler e gravar dados entre estruturas de dados na memória e diferentes formatos: CSV e arquivos de texto, Microsoft Excel, bancos de dados SQL e o formato rápido HDF5. A utilização do Pandas junto com o Python está em uma ampla variedade de domínios acadêmicos e comerciais , incluindo Finanças, Neurociência, Economia, Estatística, Publicidade, Web Analytics e muito mais.(Pandas, 2021)

5.5 **StatsModels**

StatsModels fornece classes e funções para a estimativa de muitos modelos estatísticos diferentes, bem como para a realização de testes estatísticos e exploração de dados. Uma extensa lista de estatísticas de resultados está disponível para cada estimador. Os resultados são testados em relação aos pacotes existentes para garantir que estão corretos. (StatsModels, 2021)

5.6 **Matplotlib**

Matplotlib é um pacote de gráficos 2D usado para Python para desenvolvimento de aplicativos, scripts interativos e geração de imagens com qualidade de publicação em interfaces de usuário e sistemas operacionais, sua origem oriunda do MATLAB. O Matplotlib faz o uso intenso de NumPy.(Matplotlib, 2021)

5.7 **Seaborn**

Seaborn é uma biblioteca de visualização de dados Python baseada em matplotlib . Ele fornece uma interface de alto nível para desenhar gráficos estatísticos atraentes e informativos.(Seaborn, 2021)

5.8 Plotly

Plotly é uma biblioteca de plotagem interativa de código aberto que oferece suporte a mais de 40 tipos de gráficos exclusivos, cobrindo uma ampla variedade de casos de uso estatísticos, financeiros, geográficos, científicos e tridimensionais.(Plotly, 2021)

5.9 Scikit-learn

Scikit-learn é uma biblioteca de aprendizado de máquina de código aberto que oferece suporte ao aprendizado supervisionado e não supervisionado. Ele também fornece várias ferramentas para ajuste de modelo, pré-processamento de dados, seleção e avaliação de modelo e muitos outros utilitários.(Scikit-learn, 2021)

5.10 Streamlit

Streamlit é uma biblioteca Python de código aberto que facilita a criação e o compartilhamento aplicativos da web personalizados para aprendizado de máquina e ciência de dados. (STREAMLIT INC, 2020).

O Streamlit possibilita a criação de aplicativos elegantes para modelos de machine learning (aprendizagem de máquina) ou mesmo visualização de dados para uma simples análise exploratória de um dataset (conjunto de dados), além de possuir de forma nativa HTML e JavaScript.

5.11 HTTP

O HTTP(HyperText Transfer Protocol) funciona como um protocolo de requisição e resposta no modelo computacional cliente-servidor. Um navegador web, por exemplo, pode ser o cliente e uma aplicação em um computador que hospeda um sítio da *web* pode ser o servidor. O cliente submete uma mensagem de requisição HTTP para o servidor. O servidor, que fornece os recursos, como arquivos HTML e outros conteúdos, ou realiza outras funções de interesse do cliente, retorna uma mensagem resposta para o cliente . A resposta contém informações de estado completas sobre a requisição e pode também conter o conteúdo solicitado no corpo de sua mensagem.(Tanenbaum, Andrew S.; 2011)

6. METODOLOGIA DA PESQUISA

O presente estudo caracteriza-se como uma pesquisa exploratória com abordagem qualitativa e quantitativa, onde o embasamento teórico contou com dados quantitativos. O caráter exploratório desta pesquisa caracteriza-se por trabalhar como “universo de significações, motivos, aspirações, atitudes, crenças e valores". Esse conjunto de dados considerados qualitativos corresponde a um espaço mais profundo das relações, não podendo reduzir os processos e os fenômenos à operacionalização de variáveis (MINAYO, 2001).

De acordo com Richardson (1999), a pesquisa exploratória busca conhecer as características de um fenômeno para procurar, posteriormente, explicações das suas causas e consequências. Cervo e Bervian (2002) recomendam o estudo exploratório quando há poucos conhecimentos sobre o problema a ser pesquisado, familiarizando-se com o mesmo.

Segundo Diehl e Tatim (2004), a técnica qualitativa é própria para descrever a complexidade e a interação das variáveis de determinado problema, além de compreender os processos dinâmicos vividos por grupos sociais e possibilitar, em maior nível de profundidade, o entendimento desses processos.

7. PROJETO

A aplicação utiliza a linguagem de programação Python para realizar processos e tarefas que faz a leitura das views, dos banco de dados da instituição, e as transforma em datasets, que são utilizados para as análises realizadas pela aplicação. Esse script é chamado um vez por dia no servidor, e fica responsável por se comunicar com o banco e coletar os dados pré definidos,, gerando assim informações atualizadas para serem analisadas diariamente.

O pandas, em conjunto das demais bibliotecas do Python citadas neste trabalho, fornece ferramentas que são utilizadas para a geração de gráficos apresentados pela aplicação. O Streamlit é responsável por apresentar os dados através da aplicação para os navegadores através do protocolo HTTP.

Algumas informações são possíveis de serem extraídas através de análise humana, mas para identificar padrões a fim de prever situações de evasões, utilizamos modelos de machine learning para que possam realizar essa análise, com isso é utilizado o algoritmo *random forest*.

CONSIDERAÇÕES FINAIS

Este trabalho propôs fornecer aos gestores de instituições de ensino superior privado um sistema para análise e identificação de informações que possam ser utilizadas para elaborar e gerenciar métricas para conter a alta taxa de evasão acadêmica. Este trabalho foi realizado para que venha a ser uma ferramenta valiosa para as instituições, sendo possível utilizá-lo com qualquer sistema de gestão escolar.

Os resultados obtidos ressaltam a importância de que as instituições tenham suas gestões com base em dados gerados por seus acadêmicos, extraindo informações e gerando valor para a mesma..

Além disso, seus resultados foram apresentados para uma empresa desenvolvedora de um software de gestão acadêmica e foi considerado uma ferramenta de alto valor de mercado, aliado a um sistema de gestão.

CRONOGRAMA

Atividade	Fevereiro	Março	Abril	Maio	Junho	Julho
Desenvolvimento Teórico	**X**					
Prototipagem		**X**				
Desenvolvimento do Software			**X**	**X**	**X**	
Entrega TCC II						**X**
Defesa de Banca						**X**

REFERÊNCIAS

FOSTER, P. FAWCETT, T. **Data Science para negócios**, Alta Books Rio de Janeiro, 2012

MINAYO, M. C. S. (Org.). **Pesquisa social: teoria, método e criatividade**. Petrópolis: Vozes, 2001.

RICHARDSON, R. **Pesquisa social: métodos e técnicas**. 3ª. Ed. São Paulo: Atlas, 1999.

CERVO, A.L.; BERVIAN, P.A. **Metodologia Científica**. 5ª ed. São Paulo: Prentice Hall, 2002.

DIEHL, A.A.; TATIM, D.C. **Pesquisa em ciências sociais aplicadas: métodos e técnicas. São Paulo: Pearson Prentice Hall**, 2004.

CASTRO, N. C. L. FERRARI, G. D. **Introdução à mineração de dados: conceitos básicos algoritmos e aplicações.** São Paulo: Saraiva, 2016.

SEMBAY, J. M; **EDUCAÇÃO A DISTÂNCIA: BIBLIOTECAS DE PÓLOS DE APOIO PRESENCIAL E BIBLIOTECÁRIOS**, p. 17, Florianópolis, 2009. Disponível em <https://repositorio.ufsc.br/xmlui/bitstream/handle/123456789/92872/275857.pdf?sequence=1&isAllowed=y>. Acesso em: 23 maio 2021.

MACHADO, F. RODRIGUES, N. **Big Data: O futuro dos dados e aplicações**, Saraiva, São Paulo, 2018.

IBM (org.). **Machine Learning e a Ciência de Dados com IBM Watson**. Disponível em: https://www.ibm.com/br-pt/analytics/machine-learning> Acesso em: 25 maio 2021.

SEMESP (org.). **Evasão X informação**. Disponível em:<**https://www.semesp.org.br/wp-content/uploads/2019/05/estadao.pdf**>.Acesso em: 25 junho 2021.

HIPÓLITO, O. **País perde R$ 9 bilhões com evasão no ensino superior, diz pesquisador.** Disponível em <http://g1.globo.com/educacao/noticia/2011/02/pais-perde-r-9-bilhoes-com-evasao-no-ensino-superior-diz-pesquisador.html> Acesso em 10 de junho de 2021

CRISP. (org) **CRISP-DM 1.0.**
Disponível em: <https://www.the-modeling-agency.com/crisp-dm.pdf>. Acesso em 05 maio 2021.

FAYYAD, U.; SHAPIRO, G. P.; SMYTH, P. **From Data Mining to Knowledge Discovery: An Overview**. Menlo Park, CA: AAAI Press/The MIT Press, 1996.

BASSO, D. E. **Big Data, Inovações tecnológicas,** Curitiba, Contentus, 2020.

IBGE(org.). **Panorama Nacional**. Disponível em:<https://cidades.ibge.gov.br/brasil/panorama>.Acesso em: 25 maio 2021.

IBGE(org.). **CEMPRE 2018: Número de empresas e outras organizações recua 1,8% e de empresas maiores cresce 1,1%.** Disponível em:<https://agenciadenoticias.ibge.gov.br/agencia-sala-de-imprensa/2013-agencia-de-noticias/releases/28037-cempre-2018-numero-de-empresas-e-outras-organizacoes-recua-1-8-e-de-empresas-maiores-cresce-1-1>.Acesso em: 13 abril 2021.

Smith, Jeremy; Naylor, Robin. 2011. "Dropping Out of University: A Statistical Analysis of the Probability of Withdrawal for UK University Students". Journal of the Royal Statistical Society, 164 (2): 389-405.
Disponível em:<https://www.researchgate.net/publication/4771427_Dropping_out_of_university_A_statistical_analysis_of_the_probability_of_withdrawal_for_UK_university_students> Acesso em 02 junho 2021

Bureau of Labor Statistics (org.), U.S. Department of Labor The Economics Daily, **Median weekly earnings $606 for high school dropouts, $1,559 for advanced degree holders at**
Disponível em:<https://www.bls.gov/opub/ted/2019/median-weekly-earnings-606-for-high-school-dropouts-1559-for-advanced-degree-holders.html Acesso em 10 junho 2021

BAGGI, A. S. C. **Evasão e avaliação Institucional: uma discussão bibliográfica.** Disponível em: <https://www.scielo.br/j/aval/a/RRGrQckrsd9CRGgKy4zkHXq/abstract/?lang=pt> Acesso em 15 maio 2021.

ANEXOS

Análises

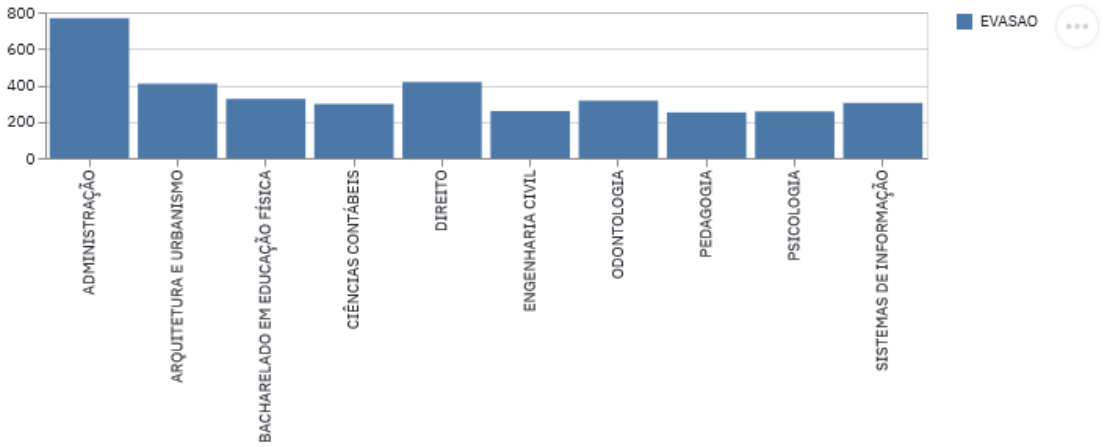
Nosso dataset tem aproximadamente 64 mil regritos e 25 colunas. Através destes dados foi realizado algumas análises apenas nas informações dos alunos que evadiram: 4875

A análise tem como ano inicial 2005 e final 2019, neste período aproximadamente 7,5% dos estudantes evadiram.

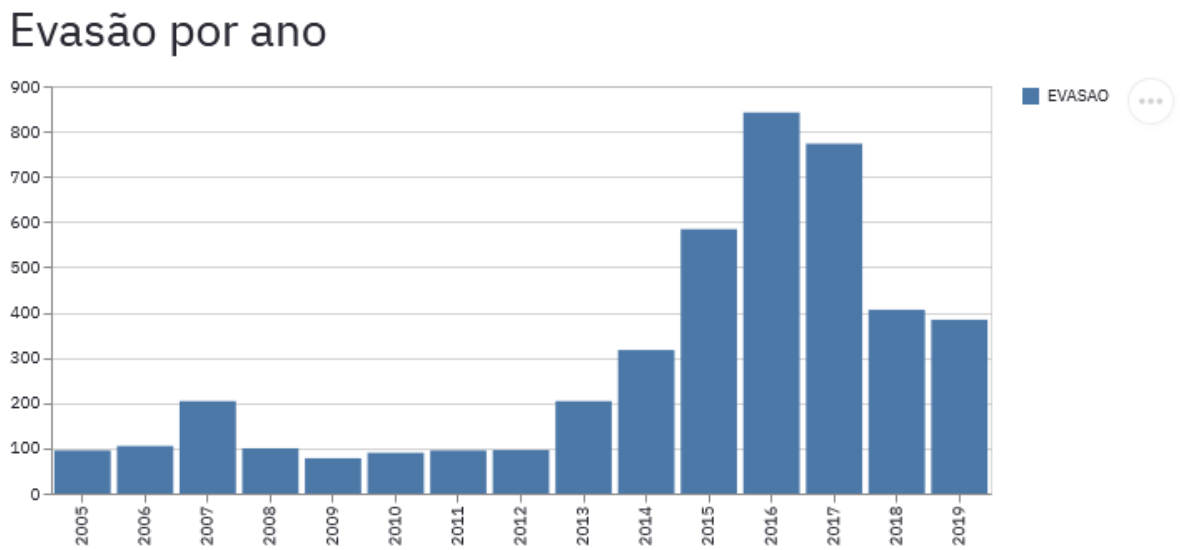
Evasão por curso

Os 10 cursos com maior número de evasão escolar:

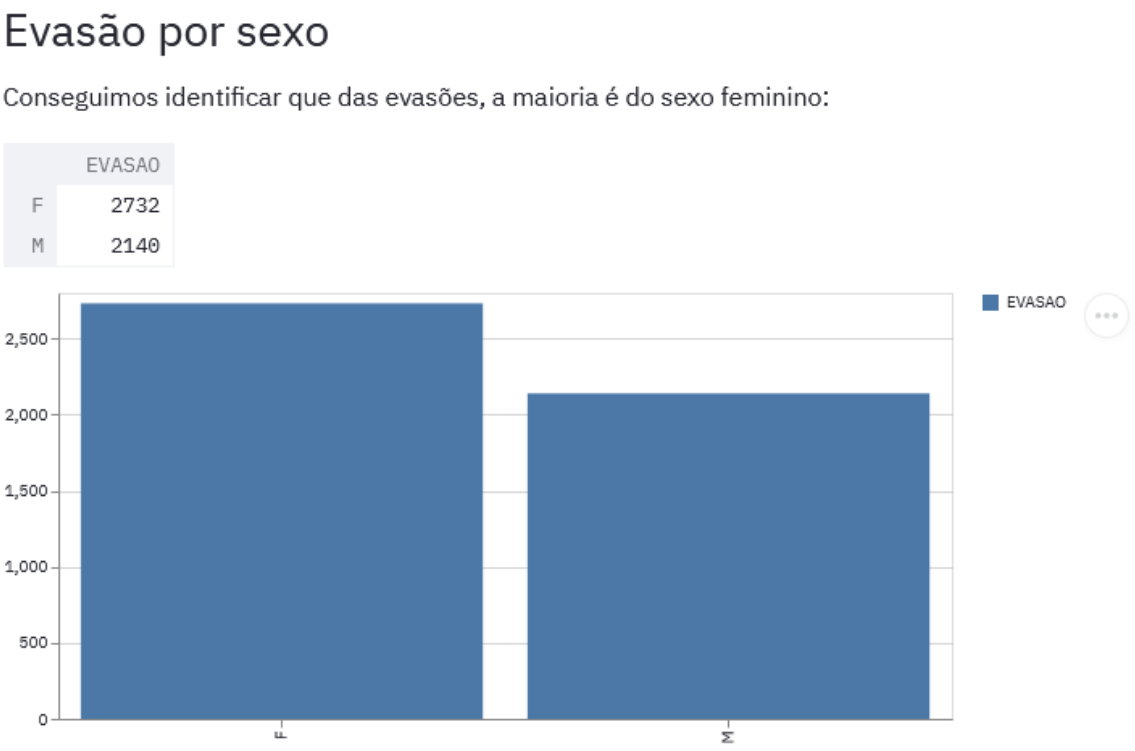
	EVASAO
ADMINISTRAÇÃO	768
DIREITO	418
ARQUITETURA E URBANISMO	409
BACHARELADO EM EDUCAÇÃO FÍSICA	325
ODONTOLOGIA	315
SISTEMAS DE INFORMAÇÃO	303
CIÊNCIAS CONTÁBEIS	297
ENGENHARIA CIVIL	258
PSICOLOGIA	256
PEDAGOGIA	251



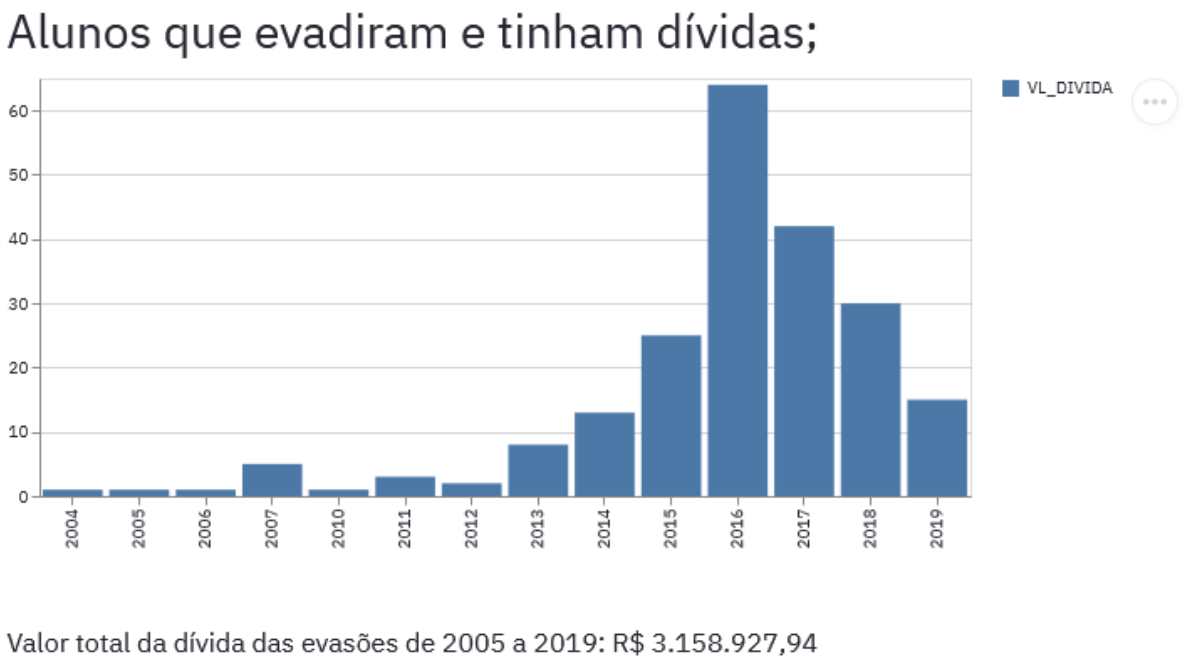
Anexo 1 – Imagem do Sistema, gráficos e informações para responder algumas perguntas gerenciais.



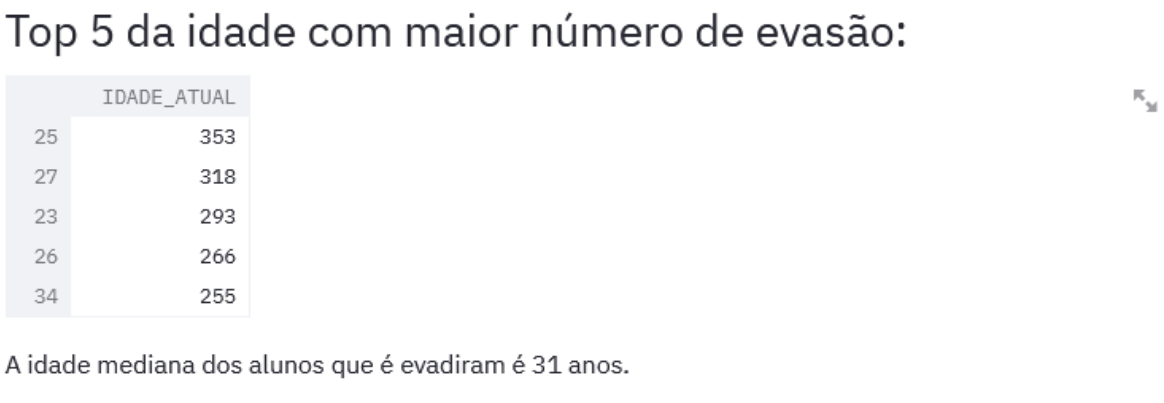
Anexo 2 – Imagem do Sistema, gráficos apresentando o número de evasões por ano.



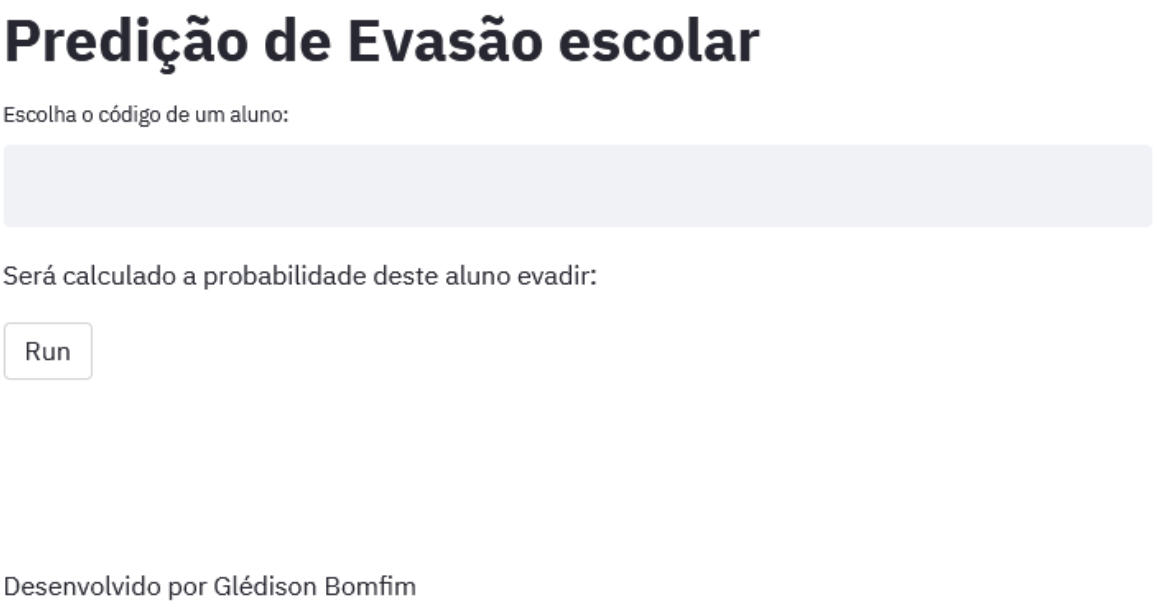
Anexo 3 – Imagem do Sistema, gráficos apresentando o número de evasões por sexo.



Anexo 4 – Imagem do Sistema, gráficos apresentando o número de evasões que possuem dívida ativa com a instituição.



Anexo 5 – Imagem do Sistema, tabela apresentando o top 5 de número de evasões por idade.



Anexo 5 – Imagem do Sistema, antes de realizar a consulta.

Anexo 6 – Imagem do Sistema, demonstrando resultado da probabilidade do estudante evadir baseado em seus dados, apresentando os dados do estudante.