UNIVERSIDADE FEDERAL DO ABC – UFABC
ENGENHARIA DE INFORMAÇÃO

FELIPE BARBOSA PEREIRA

TRABALHO DE GRADUAÇÃO EM ENGENHARIA DE INFORMAÇÃO

SANTO ANDRÉ
2024

Felipe Barbosa Pereira

TRABALHO DE GRADUAÇÃO EM ENGENHARIA DE INFORMAÇÃO

APRENDIZADO DE MÁQUINA PARA AUXILIAR O PLANEJAMENTO DE OFERTA DE DISCIPLINAS DA UNIVERSIDADE FEDERAL DO ABC

Trabalho de graduação apresentado como requisito para aprovação na disciplina Trabalho de Graduação III, do Curso de Engenharia de Informação da Universidade Federal do ABC - UFABC.

Orientador: Prof. Dr. Murilo Bellezoni Loiola

SANTO ANDRÉ

2024

Aos meus pais e minha irmã, por toda abdicação e luta em prol dos meus estudos. Se não fosse por vocês eu não seria capaz de estar aqui hoje.

AGRADECIMENTOS

Primeiramente gostaria de agradecer ao meu avô Gilberto, minha avó Regina e minha avó Nancy. Sei que para dar a possibilidade de meus pais estudarem, vocês enfrentaram batalhas que não foram nada fáceis. Todo o esforço de vocês, me deu a oportunidade de estar aqui hoje concretizando um sonho, e por conta disso, essa conquista também carrega o nome de vocês.

Agradeço aos meus pais, pois sei que houveram muitos sacrifícios por terem priorizado minha educação e a da minha irmã Beatriz. Vocês me inspiram a continuar indo atrás dos meus sonhos mesmo com todas as dificuldades que virão. Obrigado por serem meu porto seguro.

Agradeço a Beatriz, por sempre ter me ajudado nos momentos difíceis que passei ao longo da minha vida e por ter me mostrado que está tudo bem errar, isso faz parte do caminho. Serei eternamente grato por isso (apesar do acontecimento quando eu tive dengue).

Agradeço aos meus amigos e amigas que me ajudaram ao longo do turbilhão de emoções que se sucederam ao longo da graduação.

Agradeço também aos professores que me transmitiram conhecimento, seja ele técnico ou de vivência. Em especial, ao professor Murilo, por todo o suporte e compreensão ao longo da realização deste trabalho. Obrigado por fazerem de mim um engenheiro.

RESUMO

Este trabalho apresenta uma abordagem de aprendizado de máquina para auxiliar o planejamento de oferta de disciplinas da Universidade Federal do ABC. Os alunos enfrentam, dentro da universidade, o desafio da discrepância entre a quantidade de vagas ofertadas para uma disciplina e a real demanda, impedindo que parte dos alunos consigam realizar todas as disciplinas que necessitam, muitas vezes atrasando a graduação. Para isso, foi desenvolvido um sistema utilizando algoritmos de aprendizado de máquina e grafos para prever a demanda dos alunos. Com o objetivo de analisar o histórico de matrícula e identificar padrões, dados de matrículas anteriores foram coletados e utilizados para treinar um modelo de aprendizado de máquina do tipo árvore de decisão. Utilizando o modelo e grafos foi possível prever a demanda quadrimestral dos alunos. Assim, com a abordagem proposta é possível auxiliar a universidade a planejar melhor a oferta de disciplinas. Esse sistema demonstra o potencial do aprendizado de máquina para melhorar a acessibilidade nos cursos de graduação e diminuir a retenção de alunos na universidade. Partindo disso, os trabalhos futuros visam aprimorar a robustez e a acurácia do modelo através da incorporação de dados adicionais, como as matrículas em cada quadrimestre e fatores externos que impactam o desempenho dos alunos, tais como histórico de saúde, situação financeira e estrutura familiar.

Palavras-chave: aprendizado de máquina, grafos, análise de dados, sistemas de apoio à decisão.

LISTA DE FIGURAS

Figura 1- Exemplo de um grafo 14
Figura 2- Exemplo de um dígrafo 15
Figura 3- Exemplo de dados de treinamento 17
Figura 4- Exemplo de dados de teste 18
Figura 5- Exemplo de árvore de decisão 19
Figura 6- Exemplo de *DataFrame* utilizando a biblioteca Pandas 20
Figura 7- Diagrama de blocos das etapas do pré-processamento dos dados 25
Figura 8- Grafo das disciplinas obrigatórias da Engenharia de Informação 27
Figura 9- Matérias que um determinado aluno estaria apto para realizar 28
Figura 10- Exemplo de uma matrícula para o 3° quadrimestre de um ano 30
Figura 11- Estrutura dataset Dados da Realização de Disciplinas Obrigatórias 32
Figura 12- Formato da tabela para treinamento do modelo 34
Figura 13- Estrutura do sistema de apoio à decisão 35
Figura 14- Gráfico com a importância de cada *feature* 37
Figura 15- Número de alunos cursando uma determinada disciplina no segundo quadrimestre de 2023 de acordo com o sistema 38
Figura 16- Número de vagas necessárias para atender a demanda de uma determinada disciplina no terceiro quadrimestre de 2023 de acordo com o sistema 38

LISTA DE TABELAS

Tabela 1- Métricas para análise de qualidade do modelo 18
Tabela 2- Disciplinas obrigatórias para a Engenharia de Informação 23
Tabela 3- Dados disponíveis para cada período de matrícula do ano 30
Tabela 4- Colunas do dataset de treino 32

LISTA DE ABREVIAÇÕES E SIGLAS

UFABC- Universidade Federal do ABC

TSG- Taxa de Sucesso na Graduação

EI- Engenharia de Informação

DSSI- Divisão de Sistemas de Segurança de Informações

LGPD- Lei Geral de Proteção de Dados Pessoais

EIAR- Engenharia de Instrumentação, Automação e Robótica

TFAD- Teoria de Filas e Análise de Desempenho

RC- Redes de Computadores

SA- Sinais Aleatórios

TSSL- Transformada em Sinais e Sistemas Lineares

1Q- Primeiro quadrimestre

2Q- Segundo quadrimestre

3Q- Terceira quadrimestre

SUMÁRIO

1. INTRODUÇÃO 9
1.1- Problema de retenção nas universidades públicas brasileiras 9
1.2 Matrícula na Universidade Federal do ABC 10
1.3 Algoritmos para previsões 11
1.4- Aprendizado de máquina para matrícula em disciplinas 12
2. EMBASAMENTO TEÓRICO 13
2.1- Grafos 13
2.1.1- Conceitos básicos 13
2.1.2- Conceitos de um vértice 14
2.1.3- Dígrafos 15
2.2 Aprendizado de máquina 16
2.2.1 Árvores de decisão 18
2.3- Bibliotecas do Python 20
2.3.1- NetworkX 20
2.3.2- Pandas 20
2.3.3- Scikit-learn 20
3. DESENVOLVIMENTO 21
3.1- Estudo da organização do sistema de matrícula na universidade 21
3.2- Trabalho com os dados 22
3.2.1- Coleta dos dados 22
3.2.2- Limpeza e preparação dos dados 22
3.3- Definição da próxima matéria necessária para cada aluno 25
3.4- Algoritmo de aprendizado de máquina 28
3.5- Criação do sistema de suporte à decisão 34
4. RESULTADOS E DISCUSSÃO 35
4.1- Resultados do modelo de aprendizado de máquina 35
4.2- Resultados do sistema 37
4.3- Discussão sobre o sistema 39
5. TRABALHOS FUTUROS 40
6. REFERÊNCIAS BIBLIOGRÁFICAS 41

1. INTRODUÇÃO

Atualmente, é observado nas universidades públicas brasileiras o problema em que os alunos não se formam no prazo esperado, acarretando um prolongamento do tempo dos estudantes no ensino superior público, que por sua vez desencadeia um aumento nos gastos das universidades, sendo cada vez mais necessária a obtenção de recursos para sustentar a quantidade elevada de alunos.

As causas da retenção de alunos é um tema de debate e preocupação entre as universidades federais, como é possível perceber por meio dos canais de comunicação da Universidade Federal de Alagoas (SOARES, 2012) e da Universidade Federal Rural de Pernambuco (NAWA, 2019), que expuseram este problema de retenção em notícias dos próprios sites:

> "A retenção traz consigo problemas pedagógicos e administrativos, sendo, do ponto de vista educativo, a perda da visão do conhecimento integrado que o currículo pretende garantir e, do ponto de vista administrativo, a dificuldade de previsão de número de alunos por turmas, turmas excessivamente grandes e formação de turmas heterogêneas" (CESARINO et al., 1987).

Esta retenção, ocorre em diversas universidades federais brasileiras, e se dá por conta de diversos fatores como, por exemplo, a dificuldade de conciliar os estudos com o trabalho; aulas de matérias complexas em um curto período de tempo; há também, eventuais reprovações que ocorrem durante a graduação, seja por não conseguir acompanhar o curso ou por não ter aprendido corretamente os conteúdos vistos nos pré-requisitos. Um dos principais fatores que agravam a retenção acadêmica é a alta demanda para matérias, que faz com que os estudantes não consigam realizar as disciplinas que desejam ou até mesmo as que precisam, por não haver vagas suficientes para todos e nem espaço hábil para a alocação destes alunos. É fato que esta dificuldade de matrícula traz mais problemas quando está relacionada às matérias do começo da graduação; quando o aluno não se matricula nas disciplinas básicas, não consegue avançar ou realizar as matérias mais complexas que dependem do conhecimento prévio adquirido nas disciplinas iniciais.

Este problema se retroalimenta, crescendo gradativamente com o passar do tempo, pois cada ano há mais discentes entrando na universidade que discentes saindo com seus cursos finalizados. Dessa maneira, caso não haja uma melhoria

nestes processos, a graduação se tornará insustentável, tanto para os alunos que não irão conseguir realizar as matérias necessárias para se formar, quanto para a universidade que não irá conseguir manter cada vez mais alunos ingressando. Sendo assim, se torna cada vez mais necessário olhar para este problema e elaborar medidas para que a retenção de alunos diminua.

1.1- Matrícula na Universidade Federal do ABC

No contexto da Universidade Federal do ABC - UFABC, inicialmente há a realização do Bacharelado Interdisciplinar ou da Licenciatura Interdisciplinar. Após este período inicial o aluno pode escolher entre os diversos cursos específicos oferecidos pela universidade. Além disso, o período letivo é dividido em quadrimestres, ou seja, durante um ano há 3 quadrimestres. As matrículas para as disciplinas do próximo quadrimestre ocorrem durante o quadrimestre atual. Por exemplo, durante o primeiro quadrimestre do ano, os alunos escolhem as disciplinas que irão cursar no segundo quadrimestre, e assim sucessivamente. É importante ressaltar também que, na UFABC as disciplinas não têm pré-requisitos, apenas recomendações. Isso significa que os alunos podem se matricular em qualquer disciplina, mesmo que não tenham cursado as disciplinas recomendadas.

As disciplinas que serão ofertadas e a quantidade de vagas abertas dependem da disponibilidade de laboratórios, salas de aulas e professores. Esta oferta é previamente decidida pelos coordenadores dos bacharelados interdisciplinares ou dos cursos específicos e levam em consideração a matriz apresentada no projeto pedagógico, que separa todas as disciplinas obrigatórias para os alunos de acordo com o período da faculdade que o aluno se encontra.

Todavia, um dos problemas que surge a partir disto é que estes universitários muitas vezes não conseguem acompanhar a matriz do projeto pedagógico, seja por reprovação, por não ter conseguido realizar determinada matéria no quadrimestre ideal, ou até mesmo por fatores externos à universidade. Somente ofertar as disciplinas ideias demonstra que não há a visualização da realidade que o corpo discente apresenta.

Assim, a fim de diminuir esta dificuldade em se formar dada a retenção de alunos, mostrou-se necessária a intervenção no sistema de matrículas da UFABC, pois caso não seja feito nada, a tendência é que haja uma queda cada vez maior da

Taxa de Sucesso na Graduação (TSG), índice este que é utilizado por diversas universidades federais do Brasil e que se trata de um

> "[...] indicador calculado a partir da relação entre o número de formados num determinado período e o número de ingressantes de anos anteriores cuja previsão de formatura coincide com o período considerado. [...] Mais do que números baseados em diretrizes do MEC, eles sinalizam o desestímulo de cada estudante e, em muitos casos, levam ao abandono do curso." (UFF, 2015).

A fim de minimizar esta retenção de alunos e possíveis evasões é necessário que haja uma melhor alocação de disciplinas, que não leve em consideração somente a matriz pedagógica, mas também, a trajetória dos alunos, de maneira que sejam ofertadas quadrimestralmente a real necessidade do corpo discente, para que seja possível os alunos prosseguirem a graduação sem dificuldades.

1.2- Algoritmos para previsões

Os algoritmos utilizados para previsão utilizam dados históricos para detectar padrões e tendências, permitindo a projeção de valores futuros de variáveis de interesse. Suas aplicações são as mais diversas como previsão do tempo, projeções econômicas, detecção de fraudes, recomendação de produtos, diagnósticos médicos e quaisquer outros problemas que possam ser modelados a partir de dados. Previsões precisas fornecem uma base sólida para decisões, minimizando riscos e otimizando os resultados.

Existem dois tipos principais de algoritmos de previsão: os algoritmos de previsão tradicionais e os algoritmos de previsão utilizando aprendizado de máquina.

Os algoritmos tradicionais são aqueles que se utilizam de equações matemáticas, algumas simples e outras complexas, como no caso de modelos matemáticos utilizados para previsão do tempo, que necessitam de muitos dados e parâmetros para se obter boas estimativas.

Já os algoritmos de aprendizado de máquina, utilizam técnicas estatísticas e computacionais para identificar padrões complexos nos dados históricos e esses padrões são utilizados para realizar previsões sobre eventos futuros. Além disso, eles também são mais flexíveis pois se adaptam a uma variedade de tarefas de previsão e requerem grande quantidade de dados históricos para realizar o treinamento. Sendo assim, a principal diferença entre os dois é que no aprendizado

de máquina o algoritmo aprende o modelo a partir dos dados, sem o conhecimento prévio das equações matemáticas.

De acordo com Chapelle (CHAPELLE, 2006) existem três tipos principais de aprendizado de máquina: supervisionado, não supervisionado e semi-supervisionado. da

Cada um desses tipos de aprendizado de máquina possui vantagens e desvantagens. Sendo assim, a escolha do melhor algoritmo dependerá do contexto e da disponibilidade de informações.

1.3- Aprendizado de máquina para matrícula em disciplinas

No contexto de matrícula em disciplinas, o aprendizado de máquina pode oferecer diversas soluções para otimizar este processo como, por exemplo, a recomendação de disciplinas, otimização de horários e previsão de demanda, entre outras diversas possibilidades.

No intuito de melhorar a alocação de disciplinas, os algoritmos de previsão podem ser usados para prever a demanda por diferentes disciplinas nos próximos períodos letivos. A análise de dados históricos, como as matérias realizadas pelos alunos nos últimos anos, permite identificar tendências e padrões que podem ser usados para estimar, com boa precisão, o número de alunos que irão precisar de determinada matéria no futuro.

Por conta do problema pontuado neste trabalho ser uma situação que possui muitos desdobramentos relacionados diretamente a UFABC, durante a realização do mesmo, notou-se a falta de literatura para analisar soluções utilizadas em outras situações semelhantes para a matrícula em disciplinas nas universidades. Foram encontrados apenas artigos que abordam a temática de matrícula na universidade, especificamente, a da entrada de alunos na universidade e não a matrícula interna em disciplinas. Outros trabalhos encontrados foram sobre a análise de evasão de alunos utilizando aprendizado de máquina, como o artigo “Aprendizado de Máquina Aplicado à Análise de Evasão no Ensino Superior.” (PINHEIRO, et al., 2018), porém, este trabalho realiza uma análise da situação e não demonstra um meio de melhorar este cenário.

Isto demonstra que há uma lacuna importante no estudo de medidas preventivas que visem evitar a retenção de alunos nas universidades públicas. Sendo assim, o intuito deste trabalho é atuar de maneira sistemática a fim de

combater o aumento de retenção e evasão de alunos no ensino público superior e garantir um aprendizado contínuo para os estudantes de modo que eles consigam realizar as disciplinas necessárias.

Essas previsões permitirão que a Universidade Federal do ABC planeje adequadamente a alocação de recursos, como professores e disciplinas. É importante ressaltar que, inicialmente, será levado em consideração apenas os estudantes do curso de Engenharia de Informação na Universidade Federal do ABC cursando as matérias obrigatórias e, em um segundo momento, pretende-se ampliar este trabalho para os alunos do Bacharelado em Ciência e Tecnologia e do Bacharelado em Ciências e Humanidades, onde há a maior concentração de discentes. De início, não há pretensão na ampliação para as Licenciaturas Interdisciplinares pois, por enquanto, não há dados de matrícula suficientes, dado que as licenciaturas foram criadas recentemente. Outra função importante dessas previsões será permitir a adaptação não só das matérias obrigatórias como também compreender o interesse dos alunos em matérias de opção limitada e livres.

Um dos principais desafios é garantir que os modelos de previsão reflitam adequadamente a complexidade do processo educacional, sendo este influenciado por diversos fatores, como as tendências do mercado de trabalho, as políticas públicas e o interesse dos alunos.

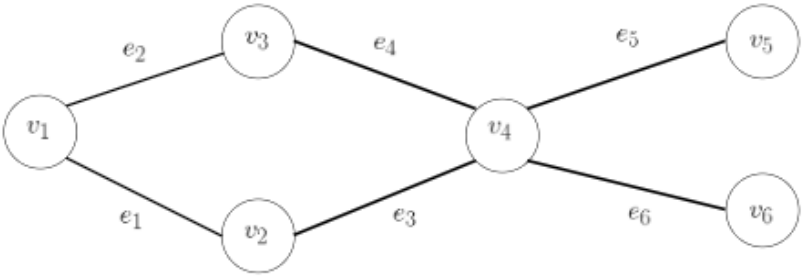
2. EMBASAMENTO TEÓRICO

2.1- Grafos

2.1.1- Conceitos básicos

Os grafos são estruturas matemáticas usadas para representar relacionamentos entre objetos. Pela terminologia usual em teoria de grafos, os objetos são denominados vértices e as relações entre eles são representadas por arestas. Conforme explicado por Bondy (BONDY, 1979), um determinado grafo “G” possui: um conjunto de vértices “V”, um conjunto de arestas “E” e uma função de incidência “Ψ”, que relaciona uma aresta a um par de vértices. Como exemplo, considere o grafo mostrado na Figura 1.

Figura 1- Exemplo de um grafo



Fonte: Autoria Própria

Neste grafo, o conjunto dos vértices é representado por

$$V = \{v_1, v_2, v_3, v_4, v_5, v_6\},$$

enquanto, o conjunto de arestas é dado por

$$E = \{e_1, e_2, e_3, e_4, e_5, e_6\}.$$

Já para a determinação da função Ψ, é necessário verificar quais os vértices que estão conectados a uma determinada aresta. Assim,

$$\begin{aligned}
\psi(e_1) &= \{v_1, v_2\} \\
\psi(e_2) &= \{v_1, v_3\} \\
\psi(e_3) &= \{v_2, v_4\} \\
\psi(e_4) &= \{v_3, v_4\} \\
\psi(e_5) &= \{v_4, v_5\} \\
\psi(e_6) &= \{v_4, v_6\}
\end{aligned}$$

.

Para um grafo ser uma árvore é necessário que ele tenha n vértices e n-1 arestas. Além disso, uma árvore não pode possuir ciclos, ou seja, o grafo não pode possuir um caminho fechado (que começa e termina no mesmo vértice).

2.1.2- Conceitos de um vértice

Um conceito importante da teoria dos grafos é o de vizinhança, pois a vizinhança de um vértice v em um grafo H é um subgrafo induzido de H constituído por todos os vértices adjacentes a v e todas as arestas ligando estes vértices.

A vizinhança de um vértice também pode ser usada para definir o conceito de grau de um vértice, que é o número de arestas que estão conectadas a ele. Sendo assim, o grau de um vértice v qualquer é igual ao tamanho da sua vizinhança.

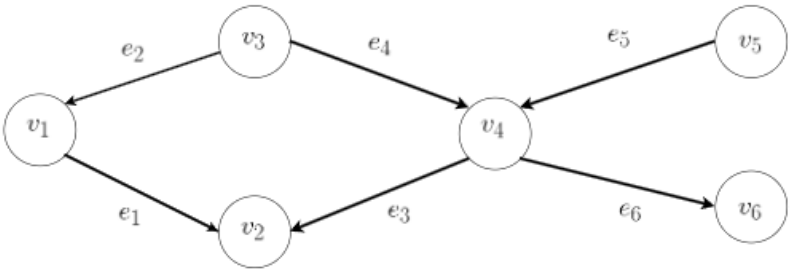
Além disso, também é possível adicionar atributos aos vértices, o que pode ampliar ainda mais o escopo de situações que pode ser representada por um grafo. Alguns exemplos relevantes disso são a definição de cor de um vértice para representar o estado em que ele se encontra ou o tipo de atributo dele. O peso dele pode representar sua importância, ou algum valor associado a ele, e também a atribuição de um texto que define uma descrição para determinado vértice ou outra informação.

2.1.3- Dígrafos

Dígrafos são grafos, cujas arestas possuem direção. Neste caso, as arestas agora são chamadas de arcos e elas não representam mais apenas pares de vértices e sim pares ordenados de vértices, para diferenciar que uma determinada aresta possui direção e um único sentido.

Abaixo, na Figura 2 é possível visualizar o grafo G mencionado anteriormente, onde suas arestas foram transformadas em arcos para que o grafo G se torne um dígrafo que chamaremos de D.

Figura 2- Exemplo de um dígrafo



Fonte: Autoria própria

No caso dos dígrafos, o conceito de vizinhança ainda vale, mas agora é designado de outra maneira. Portanto, se e=uv é um arco de um dígrafo D, então u é identificado como cabeça, enquanto o v é a cauda. Além disso, também é descrito que um arco “e” qualquer sai de “u” e entra em “v”. Assim para o dígrafo da Figura 2 em que o arco e4 sai de v3 e entra em v4.

A vizinhança em dígrafos, é divida em duas categorias: a vizinhança de entrada que são os vértices que podem alcançar o vértice v seguindo um único arco e os vizinhos de saída são os vértices que podem ser alcançados pelo vértice v

seguindo os arcos de saída. No mais, os graus de entrada representam quantos arcos entram em um determinado vértice e o grau de saída determina a quantidade de arcos que saem de um determinado vértice.

De acordo com Bondy (BONDY, 1979), nos dígrafos, o conceito de árvores também sofre uma mudança inclusive de nome, pois aqui as árvores são chamadas de arborescências e denominam a orientação de uma árvore que irá conter todos os arcos necessários para que todos os vértices sejam conectados, de modo que o grau de entrada de todos os vértices é igual a 1, com exceção da raiz desta arborescência.

2.2- Aprendizado de máquina

De acordo com Chapelle (CHAPELLE, 2006), o aprendizado de máquina é dividido em três tipos principais: supervisionado, não supervisionado e semi-supervisionado.

No aprendizado supervisionado, o algoritmo é treinado com um conjunto de dados que inclui exemplos rotulados. Os exemplos rotulados consistem em entradas (*features*) e rótulos, ou variáveis alvo, que correspondem às variáveis que o modelo está tentando prever.

Já no aprendizado não supervisionado, os dados de treino não contém rótulos. Desta forma, o algoritmo deve identificar padrões nos dados sem ser informado o que eles representam.

O aprendizado semi-supervisionado, junta os dois tipos mencionados anteriormente, pois nele o modelo é treinado com uma grande quantidade de dados não rotulados e uma pequena quantidade de dados rotulados. Deste modo o modelo aprende os padrões nos dados não rotulados a partir dos dados que foram rotulados.

No contexto do aprendizado de máquina, as *features*, também chamadas de características dos dados ou atributos, correspondem às propriedades mensuráveis de um fenômeno. Estes atributos são utilizados como instruções para o modelo, permitindo que ele aprenda a mapear determinadas entradas de atributos para saídas. Por conta das *features* darem direcionamentos para o modelo de como se comportar é importante que haja uma seleção cuidadosa das *features* do modelo.

Neste trabalho, será utilizado o modelo de aprendizado de máquina supervisionado. Como este tipo de modelo necessita de dados de treinamento para

aprender e realizar previsões precisas, é necessário definir os passos principais para a preparação desses dados.

A etapa inicial na construção de um modelo de aprendizado de máquina consiste na coleta de dados, a qual pode ser realizada através de uma ou mais fontes, seguida por um processo de curadoria para garantir a qualidade dos dados.

Embora os modelos de aprendizado de máquina geralmente tenham um melhor desempenho com conjuntos de treinamento maiores, a qualidade desses dados tem uma importância crucial. Nesse sentido, torna-se indispensável a etapa de pré-processamento, que inclui a filtragem de dados e a seleção de apenas informações relevantes para o problema em questão.

Após a formatação adequada dos dados, a próxima etapa consiste na divisão deles em dois conjuntos distintos: o conjunto de treinamento e o conjunto de testes. A proporção ideal entre treinamento e teste depende da quantidade de dados, porém uma divisão comum está em torno de 80% para treinamento e 20% para teste.

O conjunto de dados de treinamento é usado para ajustar os parâmetros do modelo escolhido de acordo com as *features* dos dados e a variável alvo. Abaixo é possível ver um exemplo de classificação binária, assim, a partir um *dataset* de treino simples, onde as variáveis x1, x2, x3 e x4 indicam as 4 *features* que esta base de dados teria.

Figura 3- Exemplo de dados de treinamento

Identificador	x1	x2	x3	x4	Variável Alvo
0	0.89	0.98	0.68	0.91	1
1	0.44	0.95	0.06	0.86	0
2	0.89	0.45	0.71	0.64	0
3	0.32	0.27	0.92	0.33	1
4	0.14	0.09	0.57	0.96	0

Fonte: Autoria própria

O segundo conjunto de dados é o conjunto de testes, composto pelas *features* relevantes para o problema em questão. A variável alvo e outras informações irrelevantes para o modelo, como o identificador, foram removidas deste conjunto.

Figura 4- Exemplo de dados de teste

x1	x2	x3	x4
0.89	0.98	0.68	0.91
0.44	0.95	0.06	0.86
0.89	0.45	0.71	0.64
0.32	0.27	0.92	0.33
0.14	0.09	0.57	0.96

Fonte: Autoria própria

A partir da criação dos dois conjuntos de dados, torna-se viável a realização do treinamento e avaliação do modelo em questão. Para avaliar a qualidade do modelo, é possível utilizar algumas métricas, conforme a tabela abaixo:

Tabela 1- Métricas para análise de qualidade do modelo

Métricas de avaliação do modelo	**Significado**
Verdadeiros Positivos (sensibilidade)	Dados que foram corretamente classificados pelo modelo como positivo.
Verdadeiros Negativos (especificidade)	Dados que foram corretamente classificados pelo modelo como negativo.
Acurácia	Dados que foram corretamente classificados pelo modelo, sejam positivos ou negativos.

Fonte: Autoria própria

Para além das métricas de desempenho mencionadas acima, a análise do impacto das *features* no modelo permite visualizar se ele está realizando correlações que não são reais. Esta análise é feita por meio do coeficiente de correlação, que mede a força da relação linear entre uma *feature* e a variável *target* e também do ganho de informação que mensura a quantidade de informação que uma feature traz para o modelo.

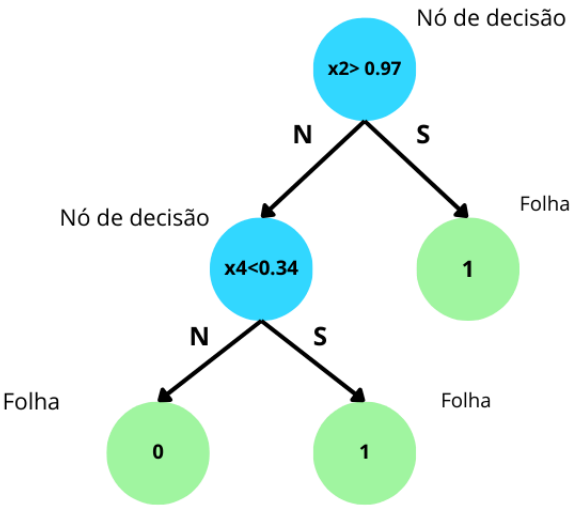
2.2.1- Árvores de decisão

A árvore de decisão é uma ferramenta de aprendizado de máquina que pode ser utilizada para diversas tarefas de classificação e regressão. Sua estrutura em

árvore permite aprender funções complexas por meio de decisões simples e interpretáveis. Existem variações do algoritmo para a construção e otimização de uma árvore.

Nas árvores de decisão, há a estrutura hierárquica em forma de árvore que permite que o algoritmo aprenda funções complexas mapeando as *features* para classes ou valores numéricos por meio de uma série de decisões simples.

Figura 5- Exemplo de árvore de decisão



Fonte: Autoria própria

Como mostrado na Figura 5, uma árvore de decisão é constituída por um conjunto de nós, que simbolizam as decisões, arcos que representam os resultados possíveis para este atributo e folhas, que representam as previsões finais. O objetivo é que a árvore criada maximize a informação obtida a cada divisão, ou seja, selecione a cada iteração a divisão ideal, que separa os dados da melhor maneira possível, criando subconjuntos com o máximo de informação sobre a classe final, para que seja necessário realizar menos divisões para chegar ao resultado final.

Cada nó verifica o valor de uma das *features* e divide o espaço de dados em dois ou mais subconjuntos. Esse processo é repetido recursivamente para cada subconjunto até que todos os dados sejam classificados.

Já para problemas de regressão, a árvore mapeia atributos para valores numéricos. Ao invés de dividir os dados em classes, cada nó estima o valor numérico com base em uma função, como a média dos valores do subconjunto, sendo as folhas uma representação dos valores de saída.

2.3- Bibliotecas do Python

2.3.1- NetworkX

A biblioteca NetworkX é utilizada para modelar grafos utilizando a linguagem de programação Python. Conforme explicado no TUTORIAL of NetworkX (NETWORKX DEVELOPERS, 2023), esta biblioteca possibilita diversas análises com um grafo, assim como a seleção dos vizinhos de determinado vértice, peso para cada vértice, análise dos diversos caminhos possíveis, criação de atributos para os vértices, entre outras funcionalidades.

2.3.2- Pandas

De acordo com a USER GUIDE da documentação do Pandas (PANDAS DEVELOPERS, 2023), que é uma biblioteca *open source* (código aberto) da linguagem de programação Python, ela é utilizada para realizar análise de dados por conta de sua praticidade e rapidez. Com a utilização desta biblioteca é possível ler uma base de dados extensa em questões de segundos e criar visualizações dos dados dentro do próprio Python, com a criação de um *DataFrame*, que é uma espécie de tabela como demonstrado na Figura 6. Além de facilitar a leitura dos dados, ela também é fundamental para a limpeza destes, por meio de comandos simples, sem haver a necessidade de analisar dado por dado para fazer a filtragem.

Figura 6- Exemplo de DataFrame utilizando a biblioteca Pandas

RA_ANONIMIZADO	ANO	PERIODO	NOME_ATIV_CURRIC	NOME_CURSO_DIPLOMA	SITUACAO	STATUS_MATRICULA
ALUNO_001410	2018	3	Circuitos Elétricos I	Engenharia de Informação	Aprovado	Aluno Regular
ALUNO_001451	2018	3	Circuitos Elétricos I	Engenharia de Informação	Aprovado	Aluno Regular
ALUNO_003242	2020	1	Circuitos Elétricos I	Engenharia de Informação	Aprovado	Aluno Regular
ALUNO_004696	2018	3	Circuitos Elétricos I	Engenharia de Informação	Aprovado	Aluno Regular
ALUNO_005565	2021	3	Circuitos Elétricos I	Engenharia de Informação	Aprovado	Aluno Regular

Fonte: Autoria própria

2.3.3- Scikit-learn

O Scikit-Learn é uma biblioteca em Python amplamente utilizada para modelagem de dados. Além disso, ela também é utilizada para a criação de modelos

de aprendizado de máquina, separação entre os dados de treinamento e dados de teste. Além disso, por meio desta biblioteca, é possível calcular facilmente algumas métricas do modelo como sensibilidade, especificidade e acurácia.

3. DESENVOLVIMENTO

O intuito do sistema de análise de matrículas criado neste trabalho é fornecer uma visualização da necessidade do corpo discente mais próxima da realidade. Pois como mencionado anteriormente, uma das causas da retenção de alunos é o problema da quantidade de vagas ofertadas para uma disciplina não ser o suficiente para a real demanda.

O sistema explicado ao longo do desenvolvimento utiliza um grafo para representar as dependências entre as disciplinas e mostrar as próximas matérias que um aluno deve realizar. E conta com um algoritmo de aprendizado de máquina para prever a partir das matérias que os alunos estão realizando quais eles irão ser aprovados ou não e, novamente, um grafo para identificar as próximas disciplinas que os alunos devem cursar.

Os passos de programação para este item estão disponíveis no GitHub, um site onde as pessoas podem acompanhar mudanças e compartilhar o código. Para visualizar os códigos dos processos listados abaixo com comentários explicativos é possível acessar o repositório através do link: https://github.com/lipebpereira/TG-UFABC

3.1- Estudo da organização do sistema de matrícula na universidade

Considerando as particularidades relacionadas à UFABC, mostrou-se necessário analisar as matrículas nas disciplinas pela perspectiva administrativa.

Por conta dessa complexidade, tornou-se necessário o contato com um profissional com expertise específica na área, para garantir o correto entendimento dos procedimentos relacionados à matrícula. Sendo assim, foram escolhidos os professores Prof. Dr. Kenji Nose Filho, coordenador do curso de Engenharia de Informação na UFABC entre os anos de 2021 e 2023 e o Prof. Dr. Murilo Bellezoni Loiola, que atuou como coordenador entre 2015 e 2017. Após as conversas, foi possível compreender o sistema de matrículas pela ótica dos coordenadores.

Durante estas conversas, foi informado que o período de matrículas funciona

da seguinte forma: antes do início de um quadrimestre, o coordenador do curso avalia quais disciplinas serão ofertadas e quais professores irão ministrar cada uma delas. Também foi destacado o desafio de identificar com precisão as demandas dos alunos a cada quadrimestre. Na tentativa de solucionar esse problema, em algumas situações anteriores, foi utilizada uma enquete entre os alunos para verificar quais matérias cada um estava precisando. De fato, após estes diálogos, a automatização deste processo mostrou-se necessária.

Dado o tempo limitado deste trabalho, foi escolhido como foco apenas o curso de Engenharia de Informação (EI) da UFABC. Porém, ao fazer esta análise para um curso, é possível utilizar a mesma estrutura para a criação de um planejamento para os outros cursos da UFABC.

3.2- Trabalho com os dados

3.2.1- Coleta dos dados

Com a finalidade de reunir os dados adequados para o desenvolvimento deste trabalho, foi necessário obter o histórico escolar dos alunos da universidade, para que fosse possível analisar a trajetória dos alunos ao longo da graduação. Assim, foi procurada a DSSI (Divisão de Sistemas de Segurança de Informações) da universidade, que disponibilizou os dados com o registro acadêmico dos alunos anonimizados, por ser uma informação sensível, para que a LGPD (Lei Geral de Proteção de Dados Pessoais) fosse respeitada. Os dados fornecidos abrangem informações da realização de disciplinas entre o primeiro quadrimestre de 2009 e o segundo quadrimestre de 2023.

3.2.2- Limpeza e preparação dos dados

Com o objetivo de tornar a base de dados mais precisa, foi necessário iniciar um processo de limpeza nestes dados, visando extrair apenas as informações relevantes para a realização do trabalho. A manipulação dos dados foi realizada utilizando a biblioteca Pandas da linguagem de programação Python.

Inicialmente havia 153 mil registros de disciplinas cursadas pelos alunos. Sendo assim, foi necessário definir um critério de quais matérias seriam analisadas. Optou-se por definir como foco da análise as disciplinas obrigatórias do curso de

Engenharia de Informação (EI) referentes ao Projeto Pedagógico de 2017 e não foram consideradas as disciplinas do Bacharelado de Ciência e Tecnologia e também as chamadas disciplinas comuns das Engenharias, assim, as disciplinas consideradas foram listadas abaixo.

Tabela 2- Disciplinas obrigatórias para a Engenharia de Informação

Item	Nome da disciplina
1	Fundamentos de Fotônica
2	Circuitos Elétricos I
3	Fundamentos de Eletromagnetismo Aplicado
4	Circuitos Elétricos II
5	Dispositivos Eletrônicos
6	Eletrônica Analógica Aplicada
7	Eletrônica Digital
8	Transformadas em Sinais e Sistemas Lineares
9	Princípios de Comunicação
10	Sistemas de Controle I
11	Redes de Computadores
12	Sinais Aleatórios
13	Processamento Digital de Sinais
14	Comunicação Digital
15	Teoria da Informação e Códigos
16	Ondas Eletromagnéticas Aplicadas
17	Comunicações Ópticas
18	Codificação de Sinais Multimídia
19	Sistemas Microprocessados
20	Comunicações Móveis
21	Teoria de Filas e Análise de Desempenho

Fonte: Autoria própria

Após filtrar apenas os registros de realização das disciplinas da Tabela 2, a quantidade de registros foi reduzida para 29 mil, tanto de alunos da EI ou não. Estes dados foram importantes para o treinamento do modelo de aprendizado de máquina, que ocorrerá na próxima seção. Sendo assim, foi definido um nome para esta base

de dados como “Dados da Realização de Disciplinas Obrigatórias”, a fim de facilitar a compreensão do desenvolvimento.

Juntamente com a organização dos dados, mostrou-se fundamental a utilização de alguns critérios para definir se um aluno está matriculado em EI ou não. Uma das principais dificuldades nesta etapa foi compreender quando um aluno está cursando Engenharia de Informação sem estar efetivamente matriculado no curso, pois na universidade há a possibilidade de alunos que ainda estão fazendo o Bacharelado Interdisciplinar realizem disciplinas de um curso específico sem estar efetivamente matriculado no curso. Existem ainda alguns casos de alunos que decidem cursar Engenharia de Informação enquanto estão matriculados em outros cursos, pois estão esperando abrir os editais internos para efetuar a troca de bacharelado.

Esta situação cria uma inconsistência nos dados, e possui diversas possibilidades de solução. A solução escolhida, neste caso foi não só levar em consideração os alunos que estão matriculados em EI, mas também levar em consideração os alunos que tivessem realizado ao menos 51% das disciplinas obrigatórias do curso, ou seja, ao menos 11 matérias específicas, levando em conta que há 21 matérias obrigatórias do curso específico (excluindo as disciplinas do Bacharelado de Ciência e Tecnologia e as chamadas disciplinas comuns das Engenharias). Com base neste critério, o sistema considera alunos que tenham cursado ao menos 11 disciplinas obrigatórias como pertencentes ao curso de Engenharia de Informação.

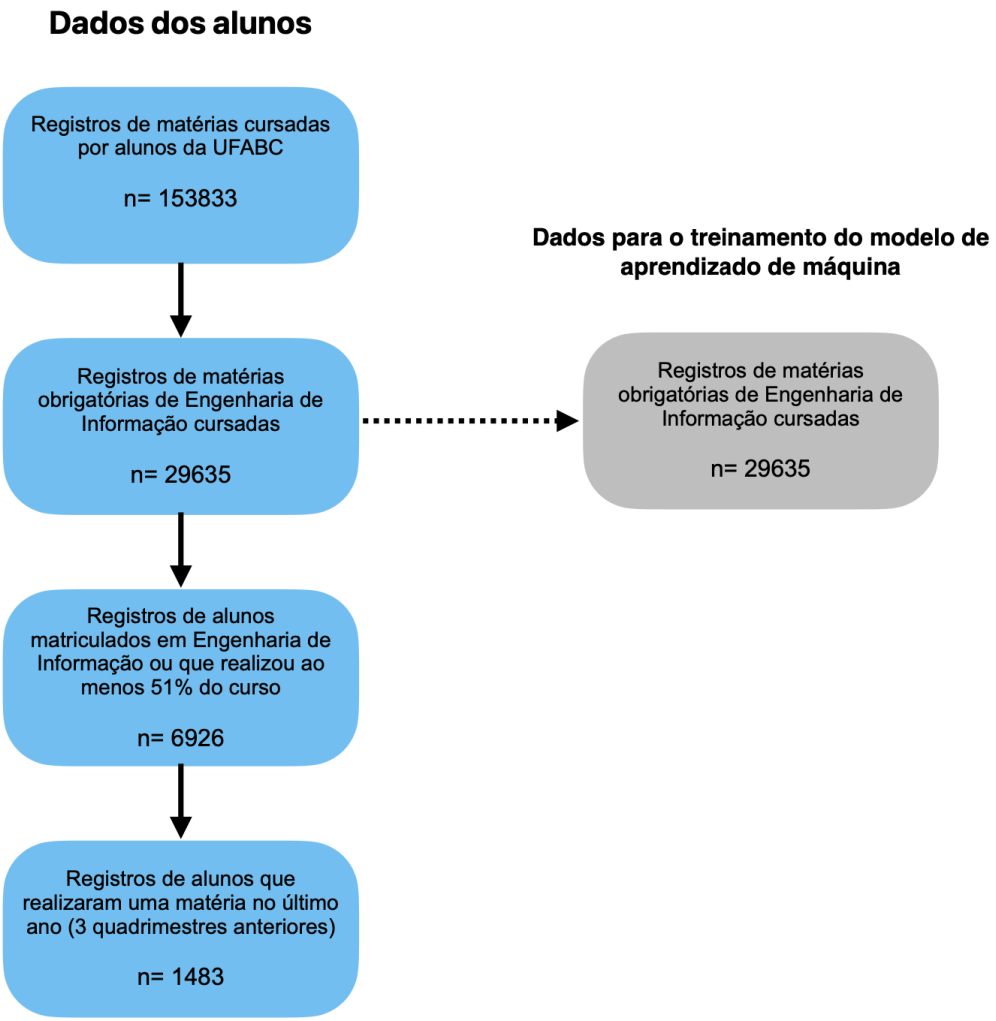
Foi utilizado a quantidade de onze disciplinas obrigatórias, para evitar possíveis conflitos em casos de engenharias que possuíssem algumas disciplinas específicas em comum, como é o caso da Engenharia de Instrumentação, Automação e Robótica (EIAR). Com este novo critério definido, a quantidade de dados foi para cerca de 7 mil registros.

Outro fator importante que teve de ser levado em consideração é o fato de que alguns alunos abandonam o curso pela metade ou trancam o curso para continuar posteriormente. Dadas estas situações, foi necessário fazer uma filtragem de alunos ativos, para isto, foi utilizado o critério de que estes alunos ativos são os alunos que realizaram ao menos uma disciplina no último ano (3 quadrimestres anteriores ao atual). Após esta filtragem, restaram por volta de 1500 registros. Note que esta filtragem não retira as disciplinas realizadas há mais de 1 ano atrás, apenas

define os alunos ativos e depois seleciona todas as disciplinas realizadas por ele, independente do ano.

Todas as etapas de pré-processamento dos dados estão resumidas na Figura 7 a seguir:

Figura 7- Diagrama de blocos das etapas do pré-processamento dos dados



Fonte: Autoria Própria

Estes dados com o histórico dos alunos que fizeram ao menos uma matéria no último ano (3 quadrimestres anteriores) é o *dataset* que será utilizado para definir a real necessidade dos alunos.

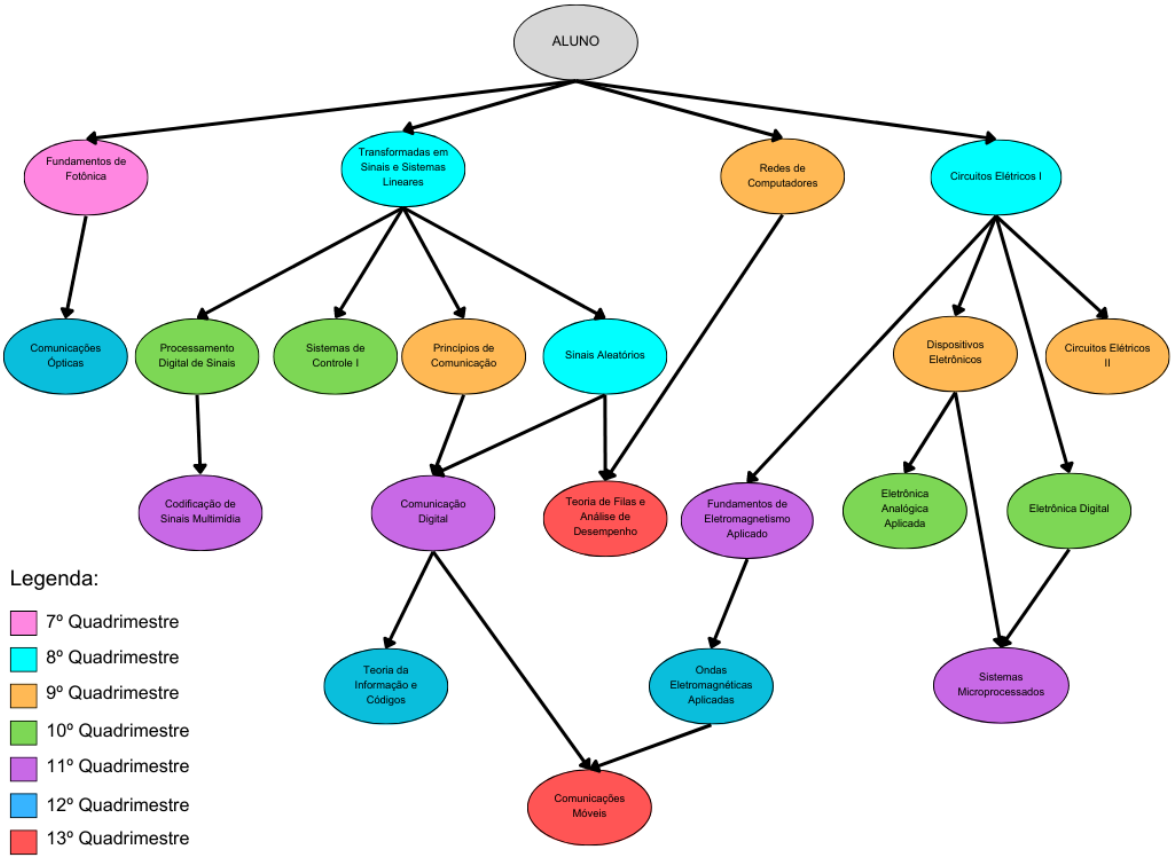
3.3- Definição da próxima matéria necessária para cada aluno

Após a filtragem de dados mencionada anteriormente, foi possível obter uma base de dados com as matérias realizadas pelos alunos ativos de EI, com um tal de 1483 registros, como mencionado anteriormente.

Neste momento, no *dataset* há as matérias já realizadas pelo aluno. Assim, é possível projetar uma visualização de quais matérias ainda precisam ser realizadas por cada aluno. Porém, quando essa análise é feita de maneira crua, apenas visualizando todas as disciplinas que os alunos ainda não realizaram, as matérias que se encontram no final do curso têm uma inflação de demanda quando comparada às matérias do começo e meio do curso. Contudo, isso não reflete a realidade, causando uma inconsistência nos dados.

Assim, mostrou-se necessário definir somente as matérias que o aluno precisa realizar no próximo quadrimestre, levando em consideração as matérias que ele já realizou. Para isso houve uma análise no projeto pedagógico do curso a fim de compreender a ordem que as disciplinas seguem. Depois deste estudo, foi definida a utilização de um tipo de grafo, a arborescência, onde os arcos representam as dependências entre as disciplinas, como demonstrado pela Figura 8, onde as cores representam o quadrimestre ideal para realização das disciplinas obrigatórias.

Figura 8- Grafo das disciplinas obrigatórias da Engenharia de Informação

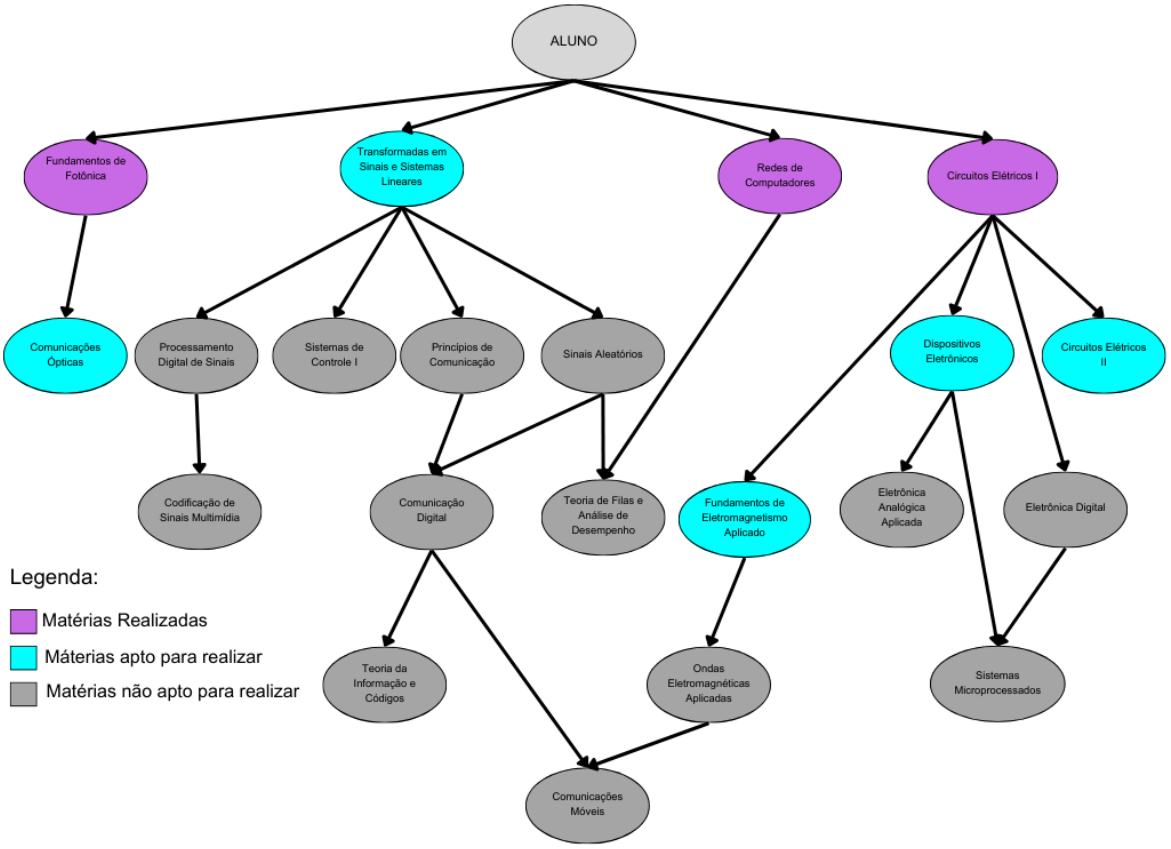


FONTE: Autoria própria.

Para a modelagem deste grafo foi utilizada a biblioteca chamada NetworkX da linguagem de programação Python. Utilizando algumas funcionalidades desta biblioteca, foi criada uma função que possui uma entrada no formato de uma lista das disciplinas que um aluno realizou e tem como saída uma lista de disciplinas que este aluno está apto a realizar.

Como exemplo, um aluno que já realizou as disciplinas Fundamentos de Fotônica, Circuitos Elétricos I e Redes de Computadores estaria apto para realizar as seguintes matérias: Comunicações Ópticas, Transformadas em Sinais e Sistemas Lineares, Fundamentos de Eletromagnetismo Aplicado, Dispositivos Eletrônicos, Circuitos Elétricos II. Estas disciplinas estão destacadas em ciano na Figura 9.

Figura 9- Matérias que um determinado aluno estaria apto para realizar



Fonte: Autoria própria

A imagem acima irá auxiliar na explicação dos conceitos que foram utilizados para definir as próximas disciplinas que um determinado aluno está apto para realizar.

É preciso salientar que para cada vértice (matéria) foram definidos dois atributos, um chamado de “visitado” (que no grafo está com a cor roxa) e o outro chamado de “candidato” (com a cor azul). O primeiro foi utilizado para denominar as disciplinas que já haviam sido realizadas pelo aluno. Já o segundo, foi utilizado para definir as matérias que o aluno estaria apto a fazer.

A análise do grafo foi feita da seguinte forma: para todos os vértices (disciplinas) que não foram visitados (não foram realizadas pelo aluno), foi feita uma análise do vértice, até a raiz do grafo. Caso todos os caminhos até a raiz do grafo estivessem preenchidos, isto é, todas as matérias recomendadas tiverem sido realizadas, então a propriedade “candidato” do vértice era alterada para “sim”, caso contrário ela era modificada para “não”. Assim, foi possível visualizar com clareza se o aluno realmente está apto para fazer determinada disciplina.

Note que na Figura 9 o aluno não está apto para a realização da disciplina Teoria de Filas e Análise de Desempenho (TFAD). Isto se dá pelo fato de o aluno ter realizado a disciplina Redes de Computadores (RC), porém não ter realizado a disciplina Sinais Aleatórios (SA) e não ter feito a disciplina anterior a esta, que é Transformada em Sinais e Sistemas Lineares (TSSL). Para este aluno estar apto para a realização da disciplina TFAD, ele deveria fazer as disciplinas TSSL e SA, pois é necessário que todos os caminhos deste vértice, até a raiz estejam preenchidos.

Com a criação deste grafo para cada aluno, foi possível uma análise adequada das próximas disciplinas a serem realizadas, de forma que seja levada em consideração se as matérias recomendadas não foram cumpridas.

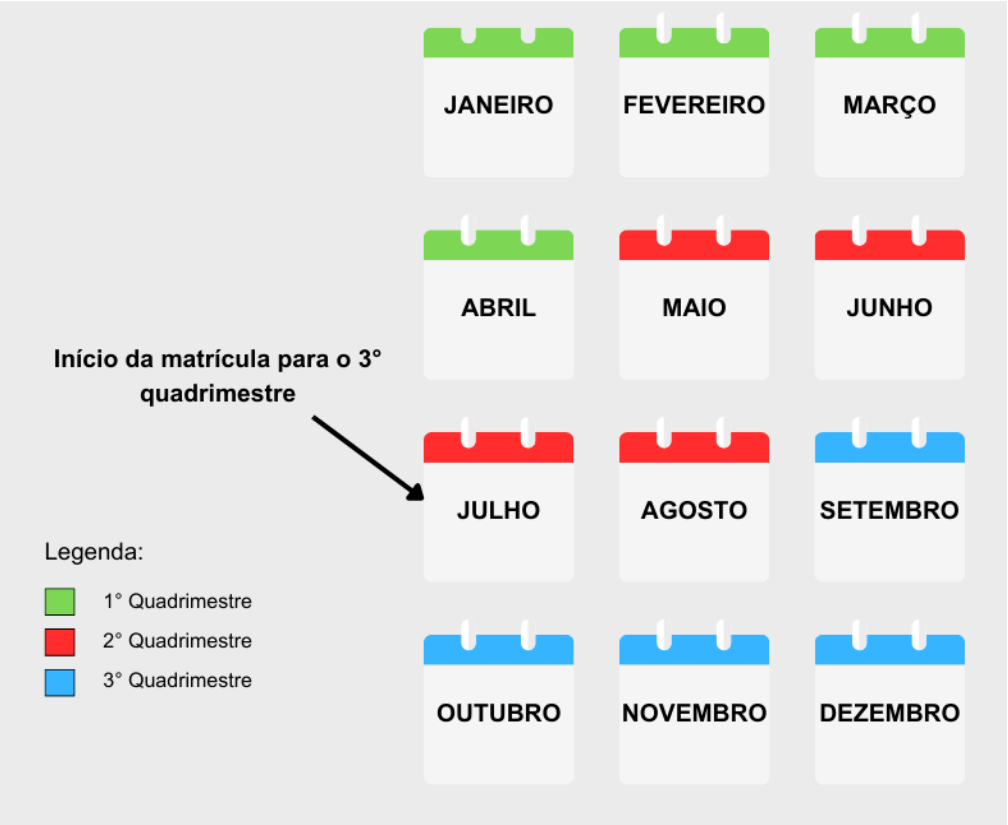
3.4- Algoritmo de aprendizado de máquina

Como visto na seção anterior, definir quais as próximas matérias que um aluno está apto para fazer é fundamental para saber a real necessidade dos alunos em um quadrimestre.

Caso as matrículas para um quadrimestre fossem realizadas ao final do quadrimestre anterior ao que está sendo analisado, apenas os grafos bastariam para se obter um retrato da realidade, pois somente seria necessário colocar as matérias que os alunos já realizaram e ver quais as próximas matérias que os alunos estão necessitando.

Todavia, conforme demonstrado na Figura 10, a matrícula para um determinado quadrimestre sempre ocorre na metade do quadrimestre precedente. Desta forma, torna-se inviável a obtenção de informações sobre as disciplinas que foram cursadas no quadrimestre anterior ao da matrícula em questão, visto que ele ainda não teria acabado.

Figura 10- Exemplo de uma matrícula para o 3° quadrimestre de um ano



Fonte: Autoria própria

Observando a figura acima percebe-se que, por exemplo, no momento da matrícula para o 3° quadrimestre (3Q) ainda não há informação sobre quais matérias foram realizadas no 2° quadrimestre (2Q), pois ele ainda está na metade. Deste modo, só há informações sobre as disciplinas realizadas até o 1° quadrimestre (1Q) do ano. Isto ocorre independente do quadrimestre que está sendo analisado, sendo que este exemplo foi utilizado a fim de facilitar o entendimento do problema.

Como é possível notar pela Tabela 3, sempre há uma defasagem de 1 quadrimestre para a obtenção atualizada de informação, o que causa uma inconsistência nos dados.

Tabela 3- Dados disponíveis para cada período de matrícula do ano

Quadrimestre da matrícula	Quando ocorre	Dados disponíveis
1° quadrimestre	Metade do 3° quadrimestre	2° quadrimestre
2° quadrimestre	Metade do 1° quadrimestre	3° quadrimestre
3° quadrimestre	Metade do 2° quadrimestre	1° quadrimestre

Fonte: Autoria própria

Para ilustrar alguns conceitos, será utilizado novamente o exemplo da matrícula no 3º quadrimestre (3Q). Por conta dessa defasagem de um quadrimestre de informação, é necessário estimar quais disciplinas os alunos estariam realizando no 2Q. Desta forma, foram colocadas as disciplinas realizadas até o 1Q para inferir quais as próximas matérias que os alunos iriam realizar. Esta estratégia permite uma aproximação das disciplinas que os alunos provavelmente estariam cursando no 2Q. Caso o fator de eventuais reprovações nas disciplinas não fosse levado em consideração, seria somente necessário colocar a estimativa de disciplinas realizadas no grafo, mostrando assim a necessidade dos estudantes para o 3Q.

Porém, eventuais reprovações devem ser levadas em consideração para se ter uma estimativa mais próxima da realidade. Por conta disso, se fez necessário a realização de algum mecanismo para realizar uma predição se um aluno irá passar ou não em uma disciplina.

Neste momento, é possível a adoção de diversas abordagens para estimar se um determinado aluno seria aprovado ou não em uma matéria. A primeira possibilidade seria uma simples multiplicação da quantidade de alunos que precisam realizar uma determinada disciplina pela porcentagem histórica de aprovação desta disciplina. A segunda abordagem pensada foi a criação de um algoritmo que leve em consideração a porcentagem de aprovação das disciplinas, a disciplina que será realizada e o histórico do aluno ao longo da graduação para saber se ele irá passar em uma determinada disciplina, ou não. Tendo em vista as opções viáveis, foi escolhida a segunda abordagem, pois ela leva em consideração não somente a disciplina, mas também o histórico escolar de cada aluno.

Para realizar a predição da aprovação ou não de um aluno em uma determinada disciplina, foi escolhido a árvore de decisão. De acordo com o embasamento teórico apresentado na Seção 2.2.1, parte fundamental para a criação de um modelo supervisionado é uma boa curadoria de dados e limpeza deles para que seja realizado um treinamento de maneira precisa.

Para realizar a criação do *dataset* de treinamento, serão utilizadas as informações do *dataset* criado anteriormente, o “Dados da Realização de Disciplinas Obrigatórias”, porém, agora será feito um filtro para levar em consideração apenas as disciplinas que têm a coluna SITUACAO como “Aprovado” ou “Reprovado”, pois

somente estes casos serão levados em consideração. Sendo assim, o *dataset* agora possui 28933 registros e está no formato demonstrado pela Figura 11.

Figura 11- Estrutura dataset Dados da Realização de Disciplinas Obrigatórias

	RA_ANONIMIZADO	NOME_ATIV_CURRIC	SITUACAO	ANO	PERIODO
0	ALUNO_000003	Transformadas em Sinais e Sistemas Lineares	Aprovado	2009	2
1	ALUNO_000003	Sistemas de Controle I	Aprovado	2010	1
2	ALUNO_000003	Circuitos Elétricos I	Aprovado	2009	1
3	ALUNO_000009	Circuitos Elétricos I	Aprovado	2014	3
4	ALUNO_000010	Transformadas em Sinais e Sistemas Lineares	Aprovado	2014	3

Fonte: Autoria própria

Dado o tempo limitado para a realização deste trabalho, adotaram-se algumas aproximações para que o modelo fosse capaz de realizar uma predição com acurácia quanto à aprovação ou reprovação de um aluno em uma disciplina específica. A seleção de atributos para o treinamento do modelo se limitou à aqueles relacionados à graduação do aluno. Os atributos (*features*) utilizados foram: o histórico do aluno durante a graduação (considerando quais matérias ele já havia feito), a disciplina que estava sendo realizada, a porcentagem de aprovação atual daquela determinada disciplina e um valor que irá demonstrar se o aluno foi aprovado ou não na realização daquela disciplina, esta que é a variável-alvo que o modelo tentará prever.

A metodologia para a criação das colunas do *dataset* de treino é descrita na Tabela 4, sendo crucial salientar que cada linha do conjunto de dados ilustrado na Figura 11 representa a matrícula de um aluno em uma disciplina específica. Sendo assim, cada linha desta base de dados irá gerar uma linha no *dataset* de treino, seguindo a seguinte estrutura:

Tabela 4- Colunas do dataset de treino

Colunas do dataset de treino	**Maneira de obter as informações destas colunas**	**Formato dos dados**

Matérias já realizadas pelo aluno	Analisar as disciplinas que foram feitas pelo aluno anteriormente à realização da disciplina que está sendo examinada.	A tabela de dados possui 21 colunas, uma para cada disciplina obrigatória e há o preenchimento de zeros e uns, onde o 0 representa as matérias que o aluno ainda não realizou e 1 as matérias que o aluno já realizou.
Disciplina que o aluno está realizando naquele quadrimestre	Analisando a coluna NOME_ATIV_CURRICULAR (coluna do arquivo disponibilizado pelo DSSI) é possível obter este dado.	Para esta coluna foi utilizado o mapeamento de disciplinas com códigos de 1 a 21, de forma sucessiva, assim, as disciplinas ficaram da seguinte forma: Fundamentos de Fotônica: 1 Circuitos Elétricos I: 2 E assim sucessivamente. Foram utilizados estes códigos para que o algoritmo de aprendizado de máquina pudesse interpretar estes valores.
Porcentagem de aprovação da disciplina	Ao olhar para a coluna SITUAÇÃO de todos os registros é possível realizar uma contagem da quantidade de aprovações e reprovações. Com isso, é necessário apenas realizar uma simples divisão entre a quantidade de aprovações e a quantidade total de registros para obter a porcentagem de aprovação de cada disciplina. É importante ressaltar que a porcentagem de aprovação de cada disciplina é coletada a partir do ano e quadrimestre (então para todos os anos e quadrimestres haverá a informação da porcentagem de aprovação de determinada disciplina de	Porcentagem de aprovação daquele quadrimestre e ano para a disciplina que está sendo realizada pelo aluno.

	maneira acumulativa).	
Indicador se o aluno foi aprovado ou não (variável-alvo)	Analisando a coluna SITUAÇÃO é possível visualizar se o aluno foi aprovado ou não.	É preenchido o valor 1 caso o aluno tenha sido aprovado na disciplina e 0 caso ele tenha sido reprovado.

Fonte: Autoria própria

A tabela que será utilizada para o treinamento do modelo, possuindo os atributos mencionados anteriormente, tem o seguinte formato:

Figura 12- Formato da tabela para treinamento do modelo

Fundamentos de Fotônica	Circuitos Elétricos I	Fundamentos de Eletromagnetismo Aplicado	Circuitos Elétricos II	Dispositivos Eletrônicos	Eletrônica Analógica Aplicada	...	Comunicações Móveis	Teoria de Filas e Análise de Desempenho	PORCENT_APROV	MATERIA	PASSOU
0	1	0	0	0	0		0	0	0.8605	8	1
0	1	0	0	0	0		0	0	0.9688	10	1
0	1	0	0	0	0		0	0	0.6146	2	1
0	1	0	0	0	0		0	0	0.7354	2	1
0	0	0	0	0	0		0	0	0.6970	8	1
...	...	...	...	...	...		...	...	...	...	...
0	0	0	0	0	0		0	0	0.7097	2	0
0	0	0	0	0	0		0	0	0.8142	6	0
0	0	0	0	0	0		0	0	0.8339	10	1
0	0	0	0	0	0		0	0	0.8304	11	0
0	0	0	0	0	0		0	0	0.8640	15	1

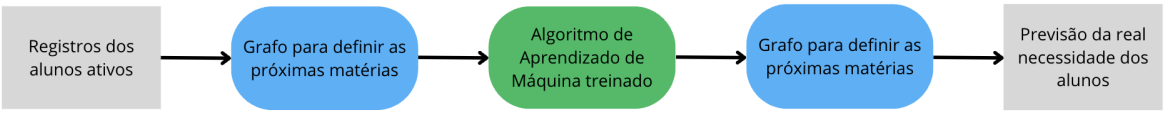
Após os dados estarem no formato desejado, foi utilizada a biblioteca do Python chamada de Scikit-Learn para realizar a separação dos dados de treinamento e dos dados de teste. Foi definido que 80% do *dataset* seria utilizado para o treinamento, que é o *dataset* que possui a variável alvo, com aproximadamente 23146 registros. 20% do *dataset* seria utilizado para testes, que tem por objetivo garantir que o modelo está ajustado para receber novos dados e definir o desempenho do modelo para estes casos, contendo 5787 registros, com dados entre primeiro quadrimestre de 2009 e o segundo quadrimestre de 2023. Após a divisão dos conjuntos de treino e teste, a biblioteca foi utilizada novamente, agora para a construção do classificador de árvore de decisão.

Este algoritmo permite a realização de predições sobre a aprovação ou reprovação de alunos em uma determinada disciplina.

3.5- Criação do sistema de suporte à decisão

Com o Grafo feito e o algoritmo de aprendizado de máquina estruturado, é possível realizar a criação de um sistema de suporte à decisão para dar ao coordenador do curso de Engenharia de Informação uma visão mais próxima da realidade acadêmica. A Figura 13 demonstra o funcionamento do sistema.

Figura 13- Estrutura do sistema de apoio à decisão



Fonte: Autoria própria

Ainda utilizando o exemplo da matrícula para o terceiro quadrimestre do ano para facilitar o entendimento, é necessário fazer uma projeção, por conta da falta de acesso aos dados das disciplinas que estariam sendo realizadas pelos alunos no segundo quadrimestre. Para isso, é utilizado o primeiro grafo para definir as próximas matérias, pois assim é possível visualizar quais matérias os alunos possivelmente estariam realizando. Novamente, esta afirmação pode ser replicada para todos os períodos letivos do ano.

Ao analisar a estrutura do sistema em questão é crucial considerar um aspecto fundamental: o segundo grafo tem como entrada os dados das disciplinas feitas no registro dos alunos ativos (o primeiro elemento do sistema, conforme ilustrado na figura 13) e as que obtiveram aprovação a partir do algoritmo de aprendizado de máquina.

Observando a Figura 13, nota-se que a partir de uma entrada com as informações dos alunos ativos do curso de Engenharia de Informação é possível obter como saída um retrato da real necessidade dos alunos. Em suma, gera-se uma matrícula justa, levando em consideração os resultados encontrados pela técnica de predição.

Outro fator relevante é que, assim como mencionado anteriormente, todo o desenvolvimento descrito neste trabalho para o curso de Engenharia de Informação da Universidade Federal do ABC pode ser replicado de maneira simples para os

outros cursos da Instituição, com a finalidade de tentar otimizar o oferecimento de disciplinas a um número maior de estudantes.

4. RESULTADOS E DISCUSSÃO

4.1- Resultados do modelo de aprendizado de máquina

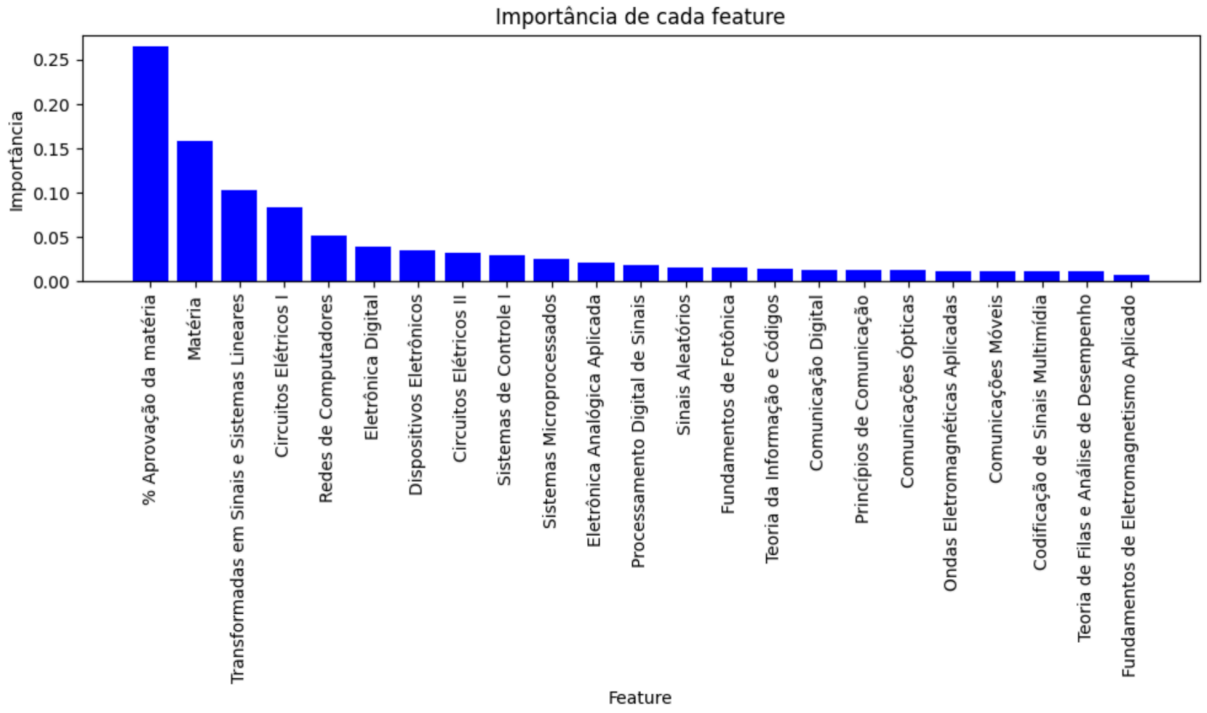
Com o intuito de definir o desempenho do algoritmo de aprendizado de máquina para a predição de aprovação ou não de alunos em disciplinas específicas, foram utilizadas métricas como acurácia, sensibilidade e especificidade, além da análise da importância de cada variável (*feature*) para o modelo.

O algoritmo apresentou resultados promissores, com uma acurácia de 80,2%, sensibilidade de 68,9% e especificidade de 84,4%. Isso indica que o modelo foi capaz de realizar uma boa predição do sucesso dos alunos na realização de disciplinas.

A análise da importância de cada *feature* foi feita por meio de uma função da biblioteca Scikit-Learn e revelou que as mais relevantes para o modelo foram:

- Porcentagem de aprovação na disciplina: Essa *feature* obteve 27% de importância. Isto faz sentido, pois ela reflete como os alunos de um modo geral se saem ao realizar aquela disciplina.
- Matéria que está sendo realizada: Essa *feature apresenta* 16% de importância, se uma matéria será mais densa ou mais simples.
- Realização da matéria Transformadas em Sinais e Sistemas Lineares: Essa *feature*, com 10% de importância, possui porcentagem maior quando comparada às outras disciplinas, pois esta é uma matéria recomendada para diversas outras disciplinas, o que justifica sua relevância.

Figura 14- Gráfico com a importância de cada *feature*

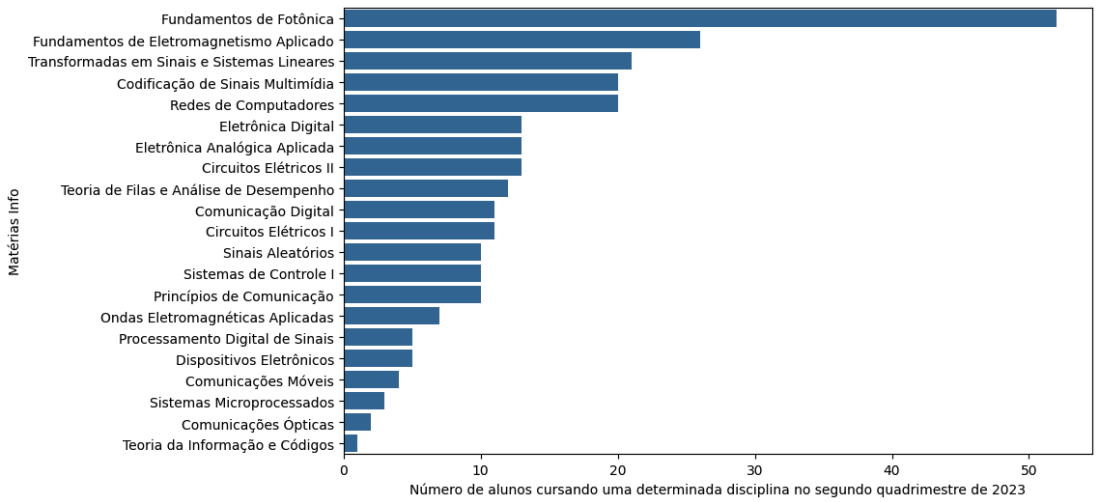


Fonte: Autoria própria

4.2- Resultados do sistema

Inicialmente, o sistema utiliza o grafo para mapear as disciplinas que os alunos estão cursando no quadrimestre atual. A figura 15 a seguir apresenta a visualização dessa informação, utilizando um gráfico de barras para mostrar a quais matérias os alunos estariam fazendo no quadrimestre anterior ao que está sendo feita a análise para a matrícula. Como mencionado ao longo do desenvolvimento, é necessário realizar essa aproximação por conta do calendário de matrículas.

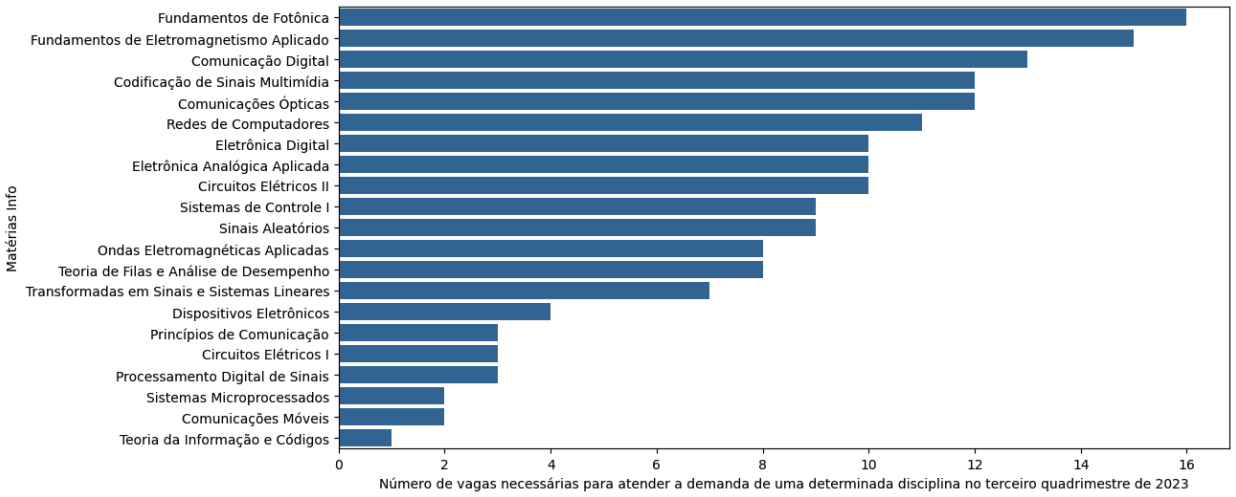
Figura 15- Número de alunos cursando uma determinada disciplina no segundo quadrimestre de 2023 de acordo com o sistema



Fonte: Autoria própria

Após a previsão de aprovação pelo algoritmo de aprendizado de máquina, o sistema utiliza novamente o grafo para identificar as próximas disciplinas que os alunos devem cursar, levando em consideração as aprovações e reprovações que foram previstas pelo modelo de aprendizado de máquina. Deste modo, a Figura 2 apresenta a visualização da quantidade de alunos que estão precisando realizar cada disciplina

Figura 16- Número de vagas necessárias para atender a demanda de uma determinada disciplina no terceiro quadrimestre de 2023 de acordo com o sistema



Fonte: Autoria própria

Ao analisar a figura acima, observa-se uma demanda maior pela disciplina "Fundamentos de Fotônica". Isso pode ser explicado pelo fato de que, durante a pandemia de COVID-19 (2020-2022), a UFABC não ofereceu turmas com aulas remotas para essa disciplina, por conta de ela possuir um grande número de aulas práticas. Essa demanda acumulada ainda persiste até o 3° quadrimestre de 2023, pois os dados do dataset vão até o 2° quadrimestre de 2023. Desta forma, a análise feita da Figura 16 retrata o último quadrimestre do ano de 2023.

Além disso, uma observação importante do resultado é que, conforme indicado pelo sistema, não há alunos precisando cursar a disciplina de Transformadas em Sinais e Sistemas Lineares. Isso ocorre porque, de acordo com o algoritmo de aprendizado de máquina, os 8 alunos (conforme demonstrado pela Figura 15) que estariam cursando esta disciplina no 2° quadrimestre de 2023 foram aprovados. Apesar de ter uma média de solicitações similar às outras disciplinas, essa matéria não possui recomendações de acordo com o grafo curricular. Assim, o fato de os estudantes estarem cursando outras disciplinas não os habilita a preencher os pré-requisitos para cursar Transformadas em Sinais e Sistemas Lineares, ao contrário do que ocorre com outras disciplinas que possuem diversas recomendações.

Deste modo, é possível notar que o sistema de apoio à decisão oferece benefícios, como:

- Melhor distribuição das vagas nas disciplinas: O coordenador dos cursos pode utilizar as informações do sistema para distribuir as vagas nas disciplinas de forma mais equilibrada, de acordo com a demanda real dos alunos.
- Redução do tempo de espera para matrícula: A previsão das matrículas permite que os alunos consigam ter a possibilidade de realizar as disciplinas quando eles precisarem.

4.3- Discussão sobre o sistema

O presente trabalho apresenta algumas limitações, em decorrência do tempo restrito para sua execução e da quantidade limitada de dados disponíveis. Para contornar essas restrições, algumas suposições foram necessárias.

Como demonstrado na Figura 10, a matrícula sempre ocorre na metade do quadrimestre anterior ao que está sendo realizada a matrícula. Por conta disso, uma

dessas limitações é a de não possuir as informações das disciplinas que estão sendo realizadas pelo aluno no 2° quadrimestre (de acordo com o exemplo da Figura 10). Desta forma, foi assumido que os alunos estariam realizando todas as disciplinas que eles estavam aptos para realizar, como no caso da primeira passagem pelo grafo. No entanto, na realidade, isso nem sempre acontece. Uma solução possível para essa questão seria obter os dados de matrícula do quadrimestre anterior ao que está sendo planejada a matrícula em questão. Assim, seria possível utilizar diretamente o algoritmo de aprendizado de máquina para analisar as aprovações e reprovações, e posteriormente o grafo, sem a necessidade de supor quais disciplinas os alunos estavam cursando naquele determinado período letivo.

Outra limitação do sistema está relacionada à falta de dados históricos sobre o desempenho dos alunos em cada disciplina. Há apenas informações sobre a nota final do aluno naquela disciplina e isso não reflete de maneira adequada o real desempenho do aluno ao longo do quadrimestre. Isso dificulta a criação de um modelo preditivo mais preciso, pois o algoritmo atual não tem acesso a informações sobre o comportamento passado dos alunos em relação às diferentes disciplinas.

Com o objetivo de auxiliar o planejamento das matrículas e oferecer suporte administrativo, o sistema implementado se provou uma ferramenta eficaz para auxiliar os coordenadores de cursos na gestão das turmas.

5. TRABALHOS FUTUROS

O projeto em questão poderia ser aprimorado através de algumas modificações. Uma delas seria a atualização do grafo, que atualmente não leva em consideração o período em que as disciplinas são oferecidas. Ao considerar o quadrimestre de cada disciplina, o grafo poderia priorizar as disciplinas iniciais do curso, as quais servem como base para diversas outras disciplinas subsequentes.

Outro ponto de melhoria diz respeito à coleta de dados. Caso houvesse a possibilidade de coletar informações no momento da matrícula, mesmo diante das evasões que eventualmente ocorram no início de cada quadrimestre, seria possível ter uma visão mais precisa das disciplinas que os alunos estariam inscritos em determinado quadrimestre. Desta forma, seria possível compreender com maior exatidão a necessidade real dos alunos.

Utilizando os dados de matrícula, seria possível levar em consideração a entrada de novos alunos no curso. No momento, o sistema não considera estes alunos, o que exige intervenção manual na disciplina Fundamentos de Fotônica para adicionar os novos ingressantes. Uma solução seria integrar os dados de matrícula ao sistema, automatizando a inclusão dos novos alunos na disciplina inicial. Isso simplificaria o processo e evitaria erros manuais.

O aprimoramento do modelo de aprendizado de máquina seria viável com a ampliação do conjunto de dados, por conta de o modelo não levar em consideração fatores externos que podem influenciar o desempenho dos alunos, como problemas pessoais, dificuldades financeiras ou mudanças no ambiente familiar. Tais fatores podem ter um impacto significativo no desempenho acadêmico dos alunos e não são considerados pelo sistema no momento. Incorporar esses fatores externos no modelo preditivo faria com que ele se tornasse mais preciso.

O sistema de apoio à decisão proposto neste trabalho tem como objetivo fornecer ao coordenador de curso uma visão mais próxima da realidade acadêmica. Através de aprimoramentos no algoritmo, a previsão do comportamento dos alunos se tornará ainda mais precisa, com dados que refletem com maior fidelidade a realidade da UFABC.

6. REFERÊNCIAS BIBLIOGRÁFICAS

MEDEIROS, C. C.; TEIXEIRA, A. A Influência Dos Atributos Da Qualidade Do Ensino Para a Retenção De Alunos De Uma Instituição De Ensino Federal. **Portal de Periódicos UFPE**, 2019. Disponível em: https://periodicos.ufpe.br/revistas/index.php/gestaoorg/article/view/231388. Acesso em ago. 2023.

IFMA. PROEN discute com os campi ações de enfrentamento à evasão e retenção. **Instituto Federal de Educação, Ciência e Tecnologia do Maranhão**, 2015. Disponível em: http://portal.ifma.edu.br/2015/10/09/proen-discute-com-os-campi-acoes-de-enfrentamento-a-evasao-e-retencao/. Acesso em: ago. 2023.

UFF. PESQUISA INÉDITA ANALISA AS CAUSAS DA RETENÇÃO DE ALUNOS DA UFF. **Universidade Federal Fluminense**, Rio de Janeiro, 29 de jun. de 2015. Disponível em: https://www.uff.br/?q=noticias/29-06-2015/pesquisa-inedita-analisa-causas-da-retencao-de-alunos-da-uff. Acesso em: ago. 2023.

NAWA, T. PREOCUPADA COM ESTUDANTES, UFRPE DISCUTE EVASÃO E RETENÇÃO NA UNIVERSIDADE. **Universidade Federal Rural de Pernambuco**, 2019. Disponível em: https://ufrpe.br/br/content/preocupada-com-estudantes-ufrpe-discute-evas%C3%A3o-e-reten%C3%A7%C3%A3o-na-universidade. Acesso em: ago. 2023.

SOARES, M. Evasão e retenção nas universidades: problemas discutidos no Forgrad 2012. **Universidade Federal de Alagoas**, 2012, editado em 2014. Disponível em: https://ufal.br/ufal/noticias/2012/12/evasao-e-retencao-nas-universidades-problemas-discutidos-no-forgrad-2012. Acesso em set. 2023.

LUZ, S. R. H.; MENDES, D. L. L .; SOARES, M. R. A. Evasão e retenção no Ensino Superior: Problemas de avaliação, currículo ou formação de professores?. **Educação & Linguagem**, 2019. Disponível em: https://www.fvj.br/revista/wp-content/uploads/2019/09/1_REdLi_2019.2.pdf. Acesso em set. 2023.

CESARINO, M. A. N.; KREMER, J. M.; DUMONT, M. M. V.; SIMÕES, E. M. S. A retenção dos alunos por infrequência, aproveitamento e trancamento de matrícula no curso de graduação em biblioteconomia da Universidade Federal de Minas Gerais. **Revista da Escola de Biblioteconomia da UFMG**, 2019. Disponível em: https://periodicos.ufmg.br/index.php/reb/article/view/36670. Acesso em set. 2023.

NETWORKX DEVELOPERS. **NetworkX: Network Analysis in Python**, 2023. Disponível em: https://networkx.org/documentation/stable/tutorial.html.

PANDAS DEVELOPERS. **Pandas Documentation**, 2023. Disponível em: https://pandas.pydata.org/docs/user_guide/index.html.

SCIKIT-LEARN DEVELOPERS. **scikit-learn User Guide**, 2023. Disponível em: https://scikit-learn.org/stable/user_guide.html.

PINHEIRO, M. A. L.; SILVA, J. C.; SOUZA, B. F. Aprendizado de Máquina Aplicado à Análise de Evasão no Ensino Superior. **Computer on the Beach**, 2018. Disponível em https://periodicos.univali.br/index.php/acotb/article/view/12810. Acesso em set. 2023.

SILVA, J. J. Uma comparação de técnicas de Aprendizado de Máquina para predição de evasão de estudantes no ensino público superior. 2022. Dissertação (Mestrado em Sistemas de Informação) - Escola de Artes, Ciências e Humanidades, University of São Paulo, São Paulo, 2022. doi:10.11606/D.100.2022.tde-23052022-092609. Acesso em out. 2023.

BONDY, J.A.; Murty, U.S.R. Graph Theory with Applications. Elsevier, New York, 1979.

CHAPELLE, O.; SCHÖLKOPF, B.; ZIEN, A. Semi-Supervised Learning. London, England: Massachusetts Institute of Technology Press., 2006. DOI: 10.1109/ICPR.2018.8546327. Disponível em https://www.molgen.mpg.de/3659531/MITPress--SemiSupervised-Learning.pdf. Acesso em dez. 2023.