

UNIVERSIDADE FEDERAL DE UBERLÂNDIA

Henrique de Moraes Segatto

**Aplicação de técnicas de projeção
multidimensional para análise do desempenho
acadêmico dos alunos de graduação da
FACOM-UFU**

Uberlândia, Brasil

2024

UNIVERSIDADE FEDERAL DE UBERLÂNDIA

Henrique de Moraes Segatto

**Aplicação de técnicas de projeção multidimensional para
análise do desempenho acadêmico dos alunos de
graduação da FCOM-UFU**

Trabalho de conclusão de curso apresentado
à Faculdade de Computação da Universidade
Federal de Uberlândia, como parte dos requi-
sitos exigidos para a obtenção título de Ba-
charel em Sistemas de Informação.

Orientador: Prof. Dr. Paulo Henrique Ribeiro Gabriel

Universidade Federal de Uberlândia – UFU

Faculdade de Computação

Bacharelado em Sistemas de Informação

Uberlândia, Brasil

2024

*Dedico a aos meus pais por me apoiarem e minhas avós, Maria Estela e Dora, que hoje
têm seus lugares entre as estrelas mais brilhantes.*

Resumo

Nos últimos anos, tem-se observado que os cursos de graduação em Tecnologia da Informação (TI) vêm demonstrando elevados índices de reprovação, acarretando uma diminuição no número de concluintes nas respectivas turmas. A análise e a compreensão das causas dessas reprovações são fundamentais tanto para os discentes que estão no processo de graduação quanto para os docentes e gestores educacionais. Técnicas de Visualização de Informação destacam-se por revelar ao usuário padrões e tendências nos dados que auxiliam na compreensão dos fenômenos observados. Essas ferramentas detêm um potencial significativo para aplicação no ambiente educacional. Contudo, a maioria dos estudos existentes utiliza dados relacionados à vida extramuros das instituições de ensino, sobretudo dados socioeconômicos referentes aos alunos. Este estudo emprega técnicas de visualização de informação e representação gráfica para investigar as representações mais adequadas dos dados coletados de alunos e analisar padrões de comportamento observados. As técnicas foram aplicadas a dados anônimos de graduandos dos cursos de Sistemas de Informação e Ciência da Computação, coletados pela Divisão de Controle Acadêmico (DICOA) da Diretoria de Administração e Controle Acadêmico (DIRAC) da Universidade Federal de Uberlândia (UFU), além de dados sintéticos criados para teste de padrões específicos. Por meio da aplicação das técnicas de visualização, é possível observar e comparar a qualidade e a correlação entre os dados completos e as técnicas de redução de dimensão, facilitando, assim, a análise e permitindo uma avaliação precisa do desempenho dos alunos em cada disciplina administrada.

Palavras-chave: Ciência de dados, Visualização de dados, Descoberta de Conhecimento em Bancos de Dados, Técnicas de Projeção Multidimensional.

Abstract

In recent years, undergraduate courses in Information Technology (IT) have shown high failure rates, resulting in a decrease in the number of graduates in their respective cohorts. Analyzing and understanding the causes of these failures are essential not only for students undergoing the graduation process but also for educators and educational administrators. Information Visualization techniques stand out for their ability to reveal patterns and trends in data, aiding in the understanding of observed phenomena. These tools hold significant potential for application in educational environments. However, most existing studies rely on data related to students' lives outside academic institutions, particularly socioeconomic data. This study employs information visualization techniques and graphical representation to investigate the most suitable representations of student data and analyze observed behavior patterns. The techniques were applied to anonymized data of undergraduates in Information Systems and Computer Science courses, collected by the Academic Control Division (DICOA) of the Administration and Academic Control Directorate (DIRAC) at the Federal University of Uberlândia (UFU), along with synthetic data created for testing specific patterns. By applying these visualization techniques, it is possible to observe and compare the quality and correlation between complete datasets and dimensionality reduction techniques, thereby facilitating analysis and enabling an accurate assessment of students' performance in each administered course.

Keywords: Data Science, Data Visualization, Knowledge Driven Discovery, Multidimensional Scaling.

Lista de ilustrações

Figura 1 – Clusterização da base de dados da flor Iris nos métodos de projeção MDS, t-SNE e g-SNE, em três repetições. As três linhas representam os três mapeamentos. - Fonte: (ZHOU; SHARPEE, 2022)	14
Figura 2 – ISOMAP aplicado sobre a base de dados Iris	15
Figura 3 – LSP aplicado sobre a base de dados Iris	16
Figura 4 – PLMP aplicado sobre a base de dados Iris	16
Figura 5 – t-SNE aplicado sobre a base de dados Iris	17
Figura 6 – Fluxograma do procedimento	19
Figura 7 – Visualização da ferramenta de visualização com filtro de seleção de área utilizado na projeção da disciplina GSI016 com o modelo LSP	20
Figura 8 – Visualização de dados da disciplina GSI002 projetada com o modelo ISOMAP.	22
Figura 9 – Visualização de dados da disciplina GSI005 projetada com o modelo ISOMAP.	22
Figura 10 – Visualização de dados da disciplina GSI002 projetada com o modelo LSP.	23
Figura 11 – Visualização de dados da disciplina GSI005 projetada com o modelo LSP.	23
Figura 12 – Visualização de dados da disciplina GSI002 projetada com o modelo PLMP.	23
Figura 13 – Visualização de dados da disciplina GSI005 projetada com o modelo PLMP.	23
Figura 14 – Visualização de dados da disciplina GSI002 projetada com o modelo t-SNE.	24
Figura 15 – Visualização de dados da disciplina GSI005 projetada com o modelo t-SNE.	24

Lista de tabelas

Tabela 1 – Atributos componentes da base	21
--	----

List of abbreviations and acronyms

Fig.	Figura
KDD	<i>Knowledge Discovery in Databases</i>
MDS	<i>Multidimensional Scaling</i>
PCA	<i>Principal Coordinates Analysis</i>
g-SNE	<i>Global t-distributed Stochastic Neighbor Embedding</i>
DICOA	Divisão de Controle Acadêmico
DIRAC	Diretoria de Administração e Controle Acadêmico
UFU	Universidade Federal de Uberlândia
FACOM	Faculdade de Computação
ISOMAP	<i>Isometric mapping</i>
LSP	<i>Least Squares Projection</i>
PLMP	<i>Part-Linear Multidimensional Projection</i>
t-SNE	<i>t-Ditributed Stochastic Neighbor Embedding</i>

Sumário

1	INTRODUÇÃO	9
1.1	Organização do trabalho	10
2	REVISÃO BIBLIOGRÁFICA	11
2.1	Ciência de Dados	11
2.2	Visualização de Informação	12
2.2.1	<i>Knowledge Discovery in Databases (KDD)</i>	12
2.2.2	Técnicas de Projeção Multidimensional	12
2.2.2.1	Técnica ISOMAP (<i>Isometric Mapping</i>)	14
2.2.2.2	Técnica LSP (<i>Least Squares Projection</i>)	15
2.2.2.3	Técnica PLMP (<i>Part-Linear Multidimensional Projection</i>)	16
2.2.2.4	Técnica t-SNE (<i>t-Distributed Stochastic Neighbor Embedding</i>)	16
2.3	Trabalhos Correlatos	17
3	DESENVOLVIMENTO	19
3.1	Ferramenta de visualização	19
3.2	Base de Dados	21
3.3	Análises	21
3.3.1	Modelo ISOMAP (<i>Isometric Mapping</i>)	22
3.3.2	Modelo LSP (<i>Least Squares Projection</i>)	23
3.3.3	Modelo PLMP (<i>Part-Linear Multidimensional Projection</i>)	23
3.3.4	Modelo t-SNE (<i>t-Distributed Stochastic Neighbor Embedding</i>)	24
4	CONCLUSÃO	25
	REFERÊNCIAS	27

1 Introdução

Nos últimos anos, os cursos de Tecnologia da Informação (TI) têm apresentado altos índices de alunos reprovados em diversas disciplinas. Com isso, o índice de formandos também foi reduzido. Estudos recentes Nesse sentido, o uso de representações gráficas de informações é uma das melhores maneiras de representar grandes aglomerados de dados visualmente, permitindo uma melhor compreensão do comportamento de uma grande base de dados, além de permitir melhores conclusões sobre esses dados (GRÉGIO et al., 2009). Para o meio em estudo não difere, pode-se identificar diversos comportamentos de dados de alunos em relação a certos períodos ou a notas de provas.

Nesse contexto, alguns estudos buscam desenvolver a análise relacionando os fatores ao redor do meio social onde o aluno se insere e suas condições econômicas ao desempenho dele nos tópicos do curso. Esta metodologia de análise utiliza fatores relacionados à vida pessoal do aluno para descrever motivos de variações no desempenho dele. Porém, para analisar a maneira como os cursos estão organizados, é preciso não só obter informações de desempenho de cada aluno, mas relacioná-las com os tópicos em que esses alunos cursaram. Utilizando essa abordagem de análise, é possível organizar dados de forma que gestores, professores e alunos possam compreender e utilizar para auxiliar em tomadas de decisão diversas.

Assim, o objetivo deste projeto é empregar um conjunto de Técnicas de Visualização de Informação na análise do desempenho acadêmico de alunos dos cursos de Sistemas de Informação e Ciência da Computação da Faculdade de Computação da Universidade Federal de Uberlândia (FACOM-UFU). Para essa análise, foram utilizados exclusivamente dados presentes nos diários e planos de ensino, de modo a considerar apenas o comportamento dos alunos nas disciplinas, bem como a estruturação dessas unidades curriculares.

Com a finalidade de alcançar o objetivo deste trabalho, os seguintes objetivos específicos foram definidos:

1. Estudar uma representação para alunos/disciplinas/cursos que facilite o processamento das técnicas de visualização de informação;
2. Prover *layouts* que comuniquem ao analista o cenário global, bem como situações específicas do contexto, permitindo a compreensão das deficiências do processo atual e guiando a criação de estratégias que mitiguem tais deficiências;
3. Prover ferramentas de interação que permitam a exploração dos *layouts* efetivamente para a extração de informações úteis para a análise.

A utilização de técnicas de Visualização de Informação em dados que representem o comportamento dos alunos em componentes curriculares pode auxiliar a busca por razões para a reprovação ou sucesso desses alunos, auxiliando os profissionais de educação a criarem políticas para mitigar esse problema. As informações coletadas também podem auxiliar os alunos com a seleção de disciplinas para cursar nos períodos seguintes. Neste estudo são usados dados de apenas dois cursos de graduação, mas é possível que eles sejam empregados para análises de outros cursos ou em outras instituições de ensino.

1.1 Organização do trabalho

O restante deste trabalho está dividido nos seguintes capítulos:

- O Capítulo 2 apresenta uma revisão sobre os conceitos básicos para entendimento do trabalho.
- O Capítulo 3 apresenta a base de dados e suas características, o processo de desenvolvimento dos testes e a análise sobre as visualizações geradas pelos testes.
- O Capítulo 4 apresenta os resultados obtidos e o ranking de quais técnicas usadas tiveram melhor efeito sobre a base.

2 Revisão Bibliográfica

Este capítulo descreve os conceitos fundamentais empregados no desenvolvimento do projeto.

2.1 Ciência de Dados

Conforme citado por (FREITAS et al., 2001), o avanço da tecnologia trouxe um grande volume de dados produzidos e obtidos de diversas fontes para o dia-a-dia, sejam eles armazenados ou em resultados de busca na Internet. Com isso, a necessidade de estudar dados armazenados tornou-se necessária em diversas áreas do conhecimento, não apenas em áreas da computação, tendo como objetivo construir informações relevantes sobre eles e os responsáveis por gerá-los. Para tal, são desenvolvidas ferramentas computacionais e algoritmos para facilitar o processo.

Os dados que armazenamos e obtemos descrevem eventos, atributos e relacionamentos de cada objeto de estudo analisado pelo usuário. Portanto, cada dado possui características específicas da informação que representam e, de acordo com estas classificações, há formas diferentes de abordar e tratar a informação com técnicas de visualização.

Existem algumas formas de classificar os dados, sejam elas pelo seu tipo de informação, tipo primitivo ou por natureza e dimensão de seu domínio. Atributos classificados pelo *tipo de informação* podem ser nominais (nomes, palavras) ou ordinais (valores numéricos), com a possibilidade de representarem uma propriedade específica de uma entidade ou relações que se constituem entre entidades (por hierarquia ou vínculo).

Já os atributos classificados devido ao *tipo dos dados*, com base nos tipos primitivos, levam em consideração se esses são inteiros, reais, alfanuméricos ou simbólicos (representando um evento, relação, etc.). Finalmente, os atributos caracterizados conforme as *dimensões de seu domínio*, sendo este unidimensional, bidimensional, tridimensional ou n-dimensional; e conforme à *natureza do domínio*, sendo ele contínuo, contínuo-discretizado ou discreto (p.ex. para um domínio bidimensional com natureza contínua, temos os dados das medidas da superfície de um terreno).

Porém, se obtivermos diversas dimensões sobre os dados, a análise desses se torna um desafio considerável, a ponto de métodos de análise de baixa dimensão não permitirem a observação de forma precisa. A abordagem que temos neste trabalho é de uso de técnicas de *visualização de informações*, ou seja, representações visuais de um conjunto de dados que permitam organizá-los e transformá-los em informação relevante, sempre a desejo do usuário que visa obter conhecimento sobre os comportamentos dos dados em questão.

2.2 Visualização de Informação

O conceito de visualização reside na capacidade de representar graficamente um conjunto de dados de forma simples e de fácil interpretação para humanos e é um campo de estudo considerável para a área de Interação Humano-Computador, que busca demonstrar diversas informações de formas onde há maior naturalidade do entendimento humano sobre seus comportamentos. Assim como relata (GRÉGIO et al., 2009), “*Frequentemente a forma mais efetiva de descrever, explorar e sumarizar um conjunto de números – mesmo um conjunto com muitos números – é ver figuras destes números*”.

Assim, entendendo as “figuras destes números”, entende-se o que é representado. Então, os gráficos diversos são desenvolvidos e planejados para análises de conjuntos de dados de acordo com a natureza deles e seus comportamentos.

2.2.1 *Knowledge Discovery in Databases* (KDD)

KDD consiste em um processo de mapear dados de baixo nível de bases de dados volumosas, através do uso de ferramentas computacionais, extraindo assim informações úteis (conhecimento) de todo o conjunto de dados que for analisado (FAYYAD; PIATETSKY-SHAPIO; SMYTH, 1996).

O processo possui diversas etapas, desde o tratamento dos dados e remoção de erros (pré-processamento) até o processo de visualização dos dados (pós-processamento), sendo este último o qual será analisado nesta pesquisa, com enfoque em técnicas de projeção dimensional.

2.2.2 Técnicas de Projeção Multidimensional

Técnicas de Projeção Multidimensional (*Multidimensional Scaling*, MDS) visam representar dados com diversos atributos e matematicamente relacioná-los para melhorar a representação dos dados em uma visualização com menores dimensões. Podemos compará-las com a projeção geométrica de uma figura de três dimensões em um plano de duas dimensões, a qual pode ou não manter as mesmas distâncias entre os pontos para formá-la em um meio de dimensões menores. A própria visualização da figura se torna outra imagem que pode manter ou acabar distorcendo a representação inicial, dependendo da necessidade de uso.

No caso das técnicas de projeção multidimensional, teremos as dimensões como os diversos atributos de uma base de dados, o que estatisticamente gera coordenadas com relação aos outros dados da base, mantendo a distância entre as coordenadas como um fator decisivo para relacionar ou diferenciar cada registro na projeção. De acordo com Borg e Groenen (2005), há quatro propósitos de uso de técnicas multidimensionais, sendo

eles:

(a) MDS como método que representa dados de (dis)similaridade em um espaço com poucas dimensões, com o objetivo de torná-las acessíveis a uma inspeção e exploração visuais; (b) como uma técnica que permite testar se e como certos critérios pelos quais se pode distinguir entre diferentes objetos de interesse são refletidos nas diferenças empíricas correspondentes desses objetos; (c) como uma abordagem data-analítica, que permite a descoberta de dimensões que fundamentam julgamentos de (dis)similaridade; (d) como um modelo psicológico que explica julgamentos de dissimilaridade em termos de uma regra que imita um tipo particular de função de distância.

Apesar de permitir a formação de grupos de dados nas visualizações, o processo de definir as coordenadas se baseia na diferença e, portanto, permite reduzir a distância de dados parecidos, a ponto de parecerem sobrepostos dependendo da técnica usada e da base de dados utilizada.

Quando é necessário observar os dados após o processo de classificação por mineração de dados, normalmente são utilizadas as técnicas de projeção com uso da classificação dada pelo processo de mineração para que seja possível diferenciar facilmente os agrupamentos de dados em um plano cartesiano, independente de quantos atributos a base de dados tenha, e determinar variância, desvio padrão, entre outras medidas sobre os dados representados.

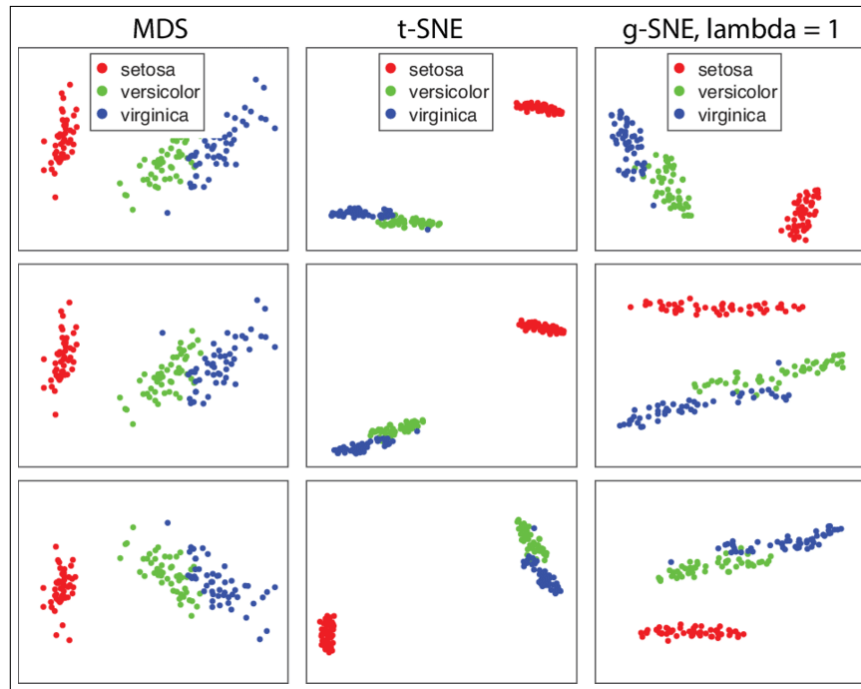


Figura 1 – Clusterização da base de dados da flor Iris nos métodos de projeção MDS, t-SNE e g-SNE, em três repetições. As três linhas representam os três mapeamentos. - Fonte: (ZHOU; SHARPEE, 2022)

De acordo com Zhou e Sharpee (2022), é possível que os modelos de projeção sejam modificados para melhor definição em certos tipos de processamento, o que permite maior customização e definição sobre a visualização dos aglomerados de dados. Como é apresentado na Figura 1, observa-se a formação das projeções de três agrupamentos feitos por mineração de dados e nota-se a diferença entre a definição distâncias através do método MDS, a qual apresenta maior variância enquanto o t-SNE apresenta maior definição de agrupamentos, porém há a sobreposição de dados, o que pode não ser ideal. Em seu projeto há a implementação de um modelo t-SNE modificado para que tenha uma visão mais precisa sobre grupos de larga escala, através de uma função de custo global o que permite manter um intermédio entre o MDS e o t-SNE, o qual ele chama de g-SNE (*Global t-distributed Stochastic Neighbor Embedding*)

2.2.2.1 Técnica ISOMAP (*Isometric Mapping*)

O modelo ISOMAP é uma técnica de redução de dimensionalidade que se destaca por preservar a estrutura geométrica das distâncias geodésicas representadas entre os pontos, com foco em capturar a topologia dos dados. Introduzido por Tenenbaum, Silva e Langford (2000), é uma ferramenta muito utilizada e aplicada em diversas áreas de estudo, como reconhecimento de padrões, processamento de imagem e bioinformática.

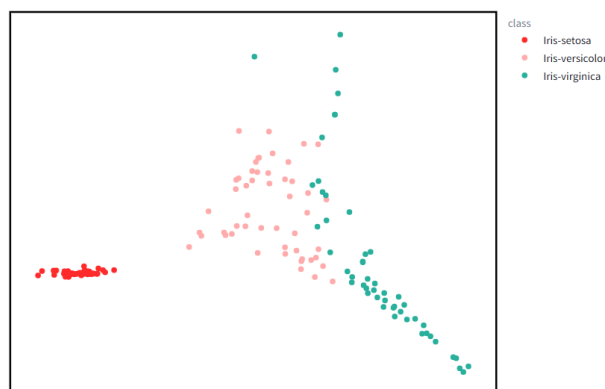


Figura 2 – ISOMAP aplicado sobre a base de dados Iris

Sabe-se que o modelo tem a capacidade de capturar a distância geodésica (a menor distância entre os dados) de forma adequada e precisa (o que outros modelos como PCA possuem limitação em casos de variedade não linear dos dados), permite uma robustez maior sobre espaços de alta dimensionalidade e a implementação mais simples e de fácil entendimento. Apesar disso, seu desempenho depende fortemente do número de vizinhos usados para construir o gráfico geodésico, o que permite, em caso de um número pequeno de vizinhos, resultar em uma má representação de distâncias globais, ou em um grande número de vizinhos, se aproxima de um método linear como MDS, perdendo a vantagem de capturar a não linearidade dos dados (BALASUBRAMANIAN; SCHWARTZ, 2002). O ISOMAP possui a tendência de representações lineares de agrupamentos, o que permite observar que quanto mais parecidos, mais linear se apresenta ao agrupamento e mais bem definido é o conjunto.

2.2.2.2 Técnica LSP (*Least Squares Projection*)

A técnica LSP foi desenvolvida com o objetivo de superar limitações de técnicas clássicas de projeção multidimensional como o Multidimensional Scaling (MDS). Utiliza a escolha de pontos de ancoragem (ou "pontos de controle"), que são usados como referência para minimizar a diferença entre as distâncias no espaço original e no espaço projetado, utilizando otimização de mínimos quadrados. De acordo com Paulovich et al. (2008), é um modelo que tem baixo custo computacional por conta de sua complexidade baixa e alta precisão, mas busca preservar a estrutura local em relação à estrutura global; por isso, pode haver distorções no todo, quando comparado com ISOMAP ou PLMP, além de ser suscetível a distorções por conta de *outliers*.

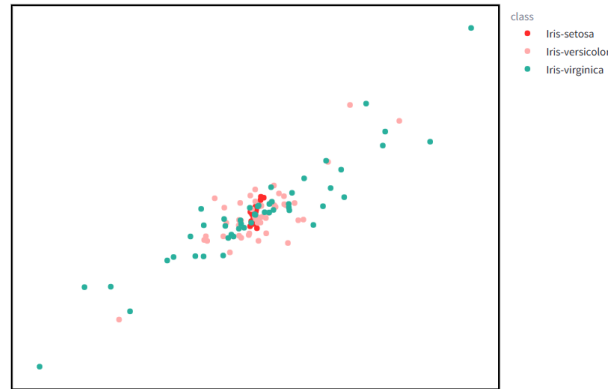


Figura 3 – LSP aplicado sobre a base de dados Iris

2.2.2.3 Técnica PLMP (*Part-Linear Multidimensional Projection*)

A técnica PLMP é uma técnica de projeção multidimensional, desenvolvida para projetar grandes bases de dados com alto desempenho e precisão. É vantajosa particularmente com bases de dados volumosas, pois utiliza apenas a informação de distância entre amostras representativas, ao invés de todos os dados, o que pode não retratar todas as nuances dos dados, comparados a métodos mais exaustivos (PAULOVICH; SILVA; NONATO, 2010).

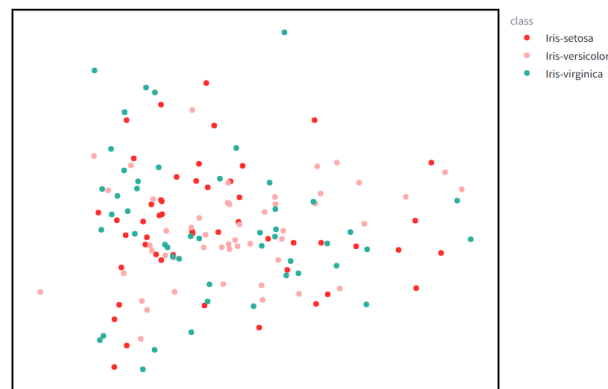


Figura 4 – PLMP aplicado sobre a base de dados Iris

2.2.2.4 Técnica t-SNE (*t-Distributed Stochastic Neighbor Embedding*)

O modelo t-SNE, criado por Maaten e Hinton (2008), é baseado em transformar as distâncias entre dois pontos de alta dimensão em probabilidades, de modo que esses pontos semelhantes na alta dimensão fiquem próximos após a redução. Demonstra alta performance com conjuntos de dados de poucas dimensões e em relação a outras técnicas não paramétricas e efetivamente revela estruturas multiescala em dados de alta dimensão.

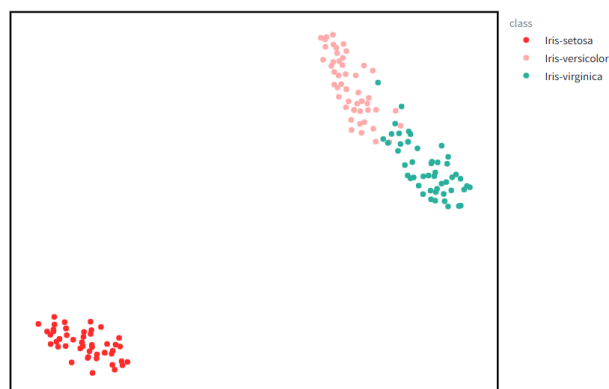


Figura 5 – t-SNE aplicado sobre a base de dados Iris

2.3 Trabalhos Correlatos

Trabalhos anteriores propuseram abordagens de visualizações focadas em analisar o comportamento acadêmico de alunos de diversos níveis de escolaridade e de formação, porém ainda há pouca documentação de testes sobre dados acadêmicos em si, apesar de alguns usos feitos para dados de alunos de instituições de ensino de outros níveis de escolaridade. Há usos diversos, desde apenas visualizar padrões até definir relações com desempenho e eventos ocorridos com o aluno.

No estudo escrito por [Silva, Magalhães Netto e Souza \(2018\)](#), é proposto o uso de ferramentas e técnicas de BI no meio educacional para monitoramento e análise do desempenho acadêmico dos alunos, através das notas das avaliações. Há certas visualizações que seriam úteis na análise, por exemplo, o gráfico de barras agrupadas de acertos e erros de questões para observação de padrões, que pode ser útil para visualizar os dados da base original, descrevendo o número de presenças, faltas e tópicos não ministrados. Apesar de ser um trabalho com enfoque apenas em alunos de matemática e em desenvolver um *dashboard* para leitura de dados de avaliações dos alunos, acredito que seria possível utilizar projeções e filtros cruzados para identificar os padrões de comportamento dos alunos.

Já no trabalho elaborado por [Borges et al. \(2018\)](#) há o uso da técnica PCA (*Principal Components Analysis*) para fazer a projeção linear e então classificação de dois *datasets* públicos que descrevem características pessoais de alunos. O estudo traz uma abordagem de testes relevante, relacionando dados dos perfis de alunos que podem contribuir na avaliação da performance do aluno e classificar mais facilmente os dados.

Finalmente, há o projeto o qual analisa o uso de Inteligência Artificial na avaliação de desempenho acadêmico de alunos do ensino médio, onde [Fernandes et al. \(2024\)](#) cita a importância de métodos como t-SNE e ISOMAP na análise de agrupamentos, permitindo relacionar dados de avaliação educacional com o desempenho de alunos. Esse tipo de abordagem é muito útil para obter informação sobre a visualização de grandes bases

educacionais.

3 Desenvolvimento

Neste capítulo, são descritos os dados usados, as técnicas de projeção multidimensional e como serão testadas e avaliadas sobre a base.

Para que as projeções possam ser comparadas, é necessário implementar um processo que permita a aplicação de técnicas de redução de dimensão nas bases de dados, além de utilizar uma representação gráfica capaz de mapear os dados detalhados na projeção, tornando o processo de análise mais intuitivo. Assim, os dados iniciais podem ser mapeados em uma visualização, possibilitando análises mais precisas sobre os fatores comportamentais de discentes e docentes.

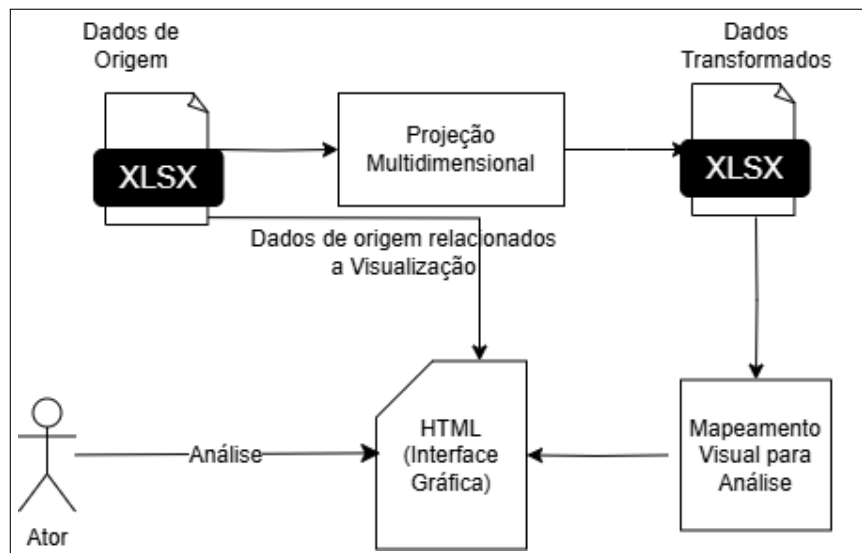


Figura 6 – Fluxograma do procedimento

O processo é composto por duas etapas. Na primeira, aplica-se algoritmos de projeção multidimensional aos dados de origem, gerando uma nova base composta por coordenadas associadas a cada aluno registrado na base inicial. Em seguida, utiliza-se uma interface intuitiva que permite relacionar os dados de presença por tópico, notas e situação final à projeção. Essa interface oferece acesso a filtros específicos, possibilitando a identificação e seleção de conjuntos formados na projeção, além da observação de comportamentos.

3.1 Ferramenta de visualização

A ferramenta de visualização foi desenvolvida utilizando a linguagem Python na versão 3.12.2 e as bibliotecas Streamlit para desenvolvimento da interface da ferramenta,

Plotly para representações gráficas e filtros interativos, Pandas para carga, manipulação e transformação de dados, Scikit-learn, Scipy e Numpy para desenvolver o *script* de execução dos modelos de projeção sobre as bases de disciplinas. Esse *script* é executado separadamente do programa, para transformar as bases de dados nas bases da projeção. Todo o código-fonte está disponível em um repositório do GitHub <<https://github.com/RikSeg/ProjetoTCC>>.

Para ilustrar, utiliza-se o processo de análise com a disciplina GSI016, com a técnica LSP, nota-se um agrupamento de 116 alunos, contendo 37 reprovados e 79 alunos aprovados, de quatro turmas diferentes, mas com comportamentos muito parecidos em relação às presenças nas aulas.

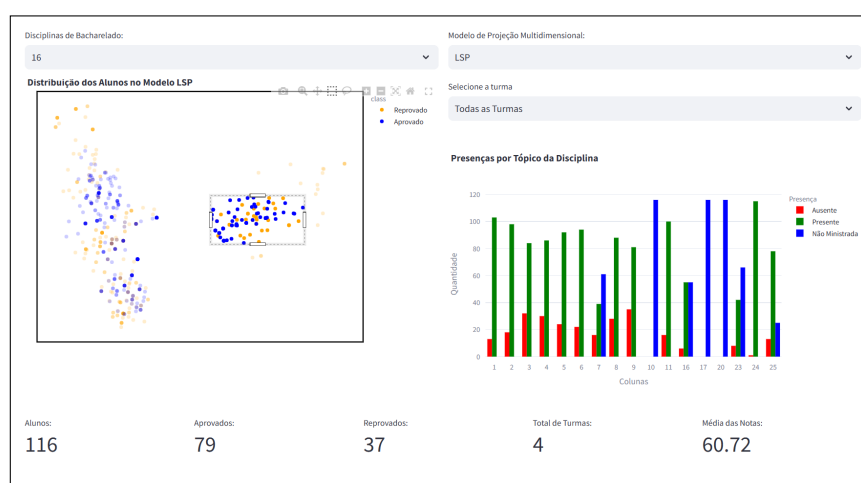


Figura 7 – Visualização da ferramenta de visualização com filtro de seleção de área utilizado na projeção da disciplina GSI016 com o modelo LSP

A Figura 7 mostra que em todas as quatro turmas, não foram ministrados os tópicos 10, 17 e 20, que provavelmente foram considerados tópicos não tão relevantes pelos docentes. Através da análise por turma, percebe-se que é mantido um número considerável de faltas no período entre os tópicos 2 e 9, porém em poucos casos há a falta dos alunos após o tópico 11 e no tópico 1. A média de notas dos alunos nesse agrupamento foi de 60,72 pontos, considerando que 68% destes foram aprovados.

Observa-se assim, o comportamento dos alunos de faltar durante os nove primeiros tópicos, mas uma presença maior nos tópicos seguintes, o que pode representar uma maior preocupação dos alunos em obter notas melhores após uma avaliação e a projeção corresponde corretamente aos dados apresentados, portanto, podemos dizer que há uma precisão considerável sobre o agrupamento dos alunos, o que pode dizer que eles tiveram situações parecidas e que afetaram diretamente o desempenho deles na avaliação da matéria.

3.2 Base de Dados

Os dados utilizados foram previamente transformados por outros pesquisadores do grupo e fornecidos pela Divisão de Controle Acadêmico (DICOA) da UFU. Foi necessário catalogar os tópicos das disciplinas da grade curricular da FACOM, de acordo com a ementa da disciplina. Foram escolhidas amostras de alunos de diversos períodos que cursaram sete disciplinas da grade curricular do curso de Bacharelado em Sistemas de Informação. A base de dados é dividida por disciplina, para que sejam projetadas individualmente por cada modelo de projeção usado neste trabalho.

A estrutura da base foi definida em dois grupos de dados. As primeiras cinco colunas são usadas para dados específicos sobre cada aluno: identificador do aluno, o identificador da turma, o código da disciplina, a nota do aluno e status de aprovação. Já as colunas restantes representam os tópicos presentes na ementa da disciplina, ou seja, tanto aulas quanto avaliações, revisão e vista de prova. Os tópicos utilizados para a visualização são apenas aulas que possuem matérias da ementa, sejam práticas ou teóricas, como o enfoque é em presença de alunos em matérias da ementa.

Campo	Exemplo de Valor
IDT_MATRICULA	1234
IDT_TURMA	5678
COD_DISCIPLINA	GS1000
VLR_MEDIA	85
DSC_SITUACAO_FINAL	Aprovado; Reprovado
Tópico 1, Tópico 2, ...	0(Ausente); 1(Presente); 2(Não Ministrada)

Tabela 1 – Atributos componentes da base

O critério que define a presença dos alunos considera que certos tópicos foram ministrados em mais de uma aula, portanto, considera-se presente o discente com presença em 70% das aulas daquele mesmo tópico. Não há registros duplicados de alunos, portanto não serão observados registros de repetentes nas análises.

3.3 Análises

Considerando que cada modelo de projeção multidimensional tem um comportamento esperado sobre os dados, de forma que se permite observar e demonstrar as diferentes maneiras que podem formar os agrupamentos, serão usadas as bases das disciplinas GSI002 e GSI005 para exemplificar os resultados esperados para cada técnica. Os gráficos de dispersão serão as visualizações escolhidas para a base, considerando a coloração das coordenadas como a classe que define qual a situação final do aluno, onde laranja equivale a "Reprovado" e azul equivale a "Aprovado".

3.3.1 Modelo ISOMAP (*Isometric Mapping*)

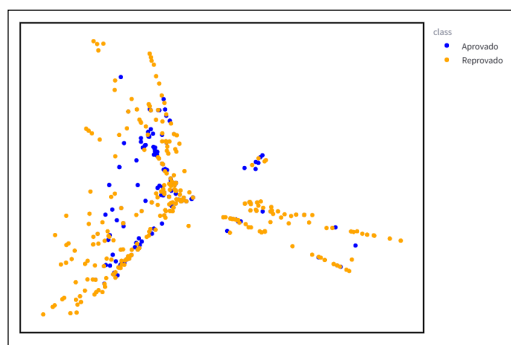


Figura 8 – Visualização de dados da disciplina GSI002 projetada com o modelo ISOMAP.

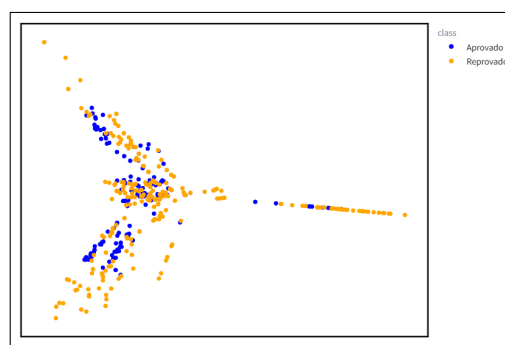


Figura 9 – Visualização de dados da disciplina GSI005 projetada com o modelo ISOMAP.

Observa-se na disciplina GSI002, representada na Figura 8 uma separação parcial entre os grupos, com uma densidade maior de alunos reprovados na extremidade direita da projeção. Esse comportamento sugere que há uma diferenciação significativa entre os alunos aprovados e reprovados, mas ainda existe alguma sobreposição entre as duas classes.

A organização dos dados também indica a presença de subgrupos onde predominam alunos aprovados, porém é um tanto difícil determinar as diferenças, principalmente dos dados no aglomerado à esquerda. A extensão à direita pode refletir padrões de faltas que estão mais associados à reprovação dos alunos.

Também podemos observar que na disciplina GSI005, representada na Figura 9 que a separação é menos evidente, com maior sobreposição entre as classes, com aglomerados mistos e algumas regiões de predominância de alunos reprovados. Nota-se a homogeneidade das presenças entre as duas classes, mas maior definição do espaçamento e de subgrupos em comparação com GSI002.

3.3.2 Modelo LSP (*Least Squares Projection*)

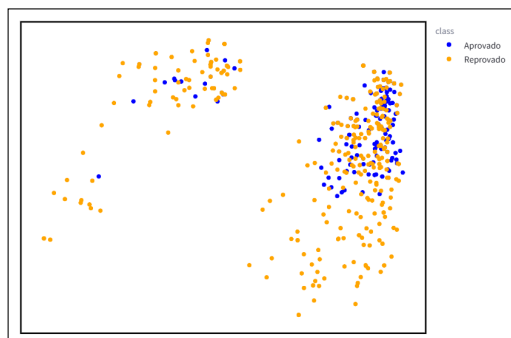


Figura 10 – Visualização de dados da disciplina GSI002 projetada com o modelo LSP.

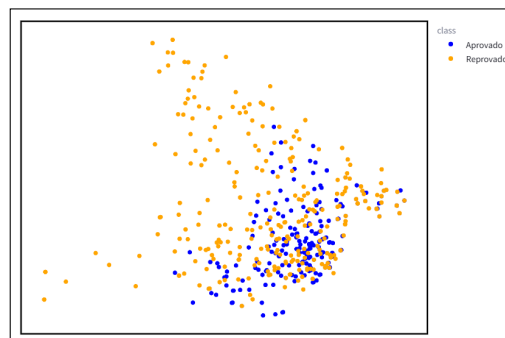


Figura 11 – Visualização de dados da disciplina GSI005 projetada com o modelo LSP.

A figura 10, uma projeção do modelo LSP da disciplina GSI002, apresentou dois principais aglomerados com uma certa separação entre as classes. Observa-se que os alunos reprovados tendem a se concentrar em um dos aglomerados mais isolados, o que pode significar costumes de presenças que vários dos alunos reprovados tiveram, enquanto alunos aprovados têm uma distribuição mais concentrada ao grupo mais à esquerda do gráfico, com distribuição mais uniforme.

A Figura 11, que representa a projeção LSP de GSI005, apresentou um agrupamento único, o que dificultou verificar padrões relacionados ao desempenho dos alunos em situações específicas, mas visivelmente definindo a classe dos alunos em dois subgrupos, apesar de ter uma concentração mais homogênea.

3.3.3 Modelo PLMP (*Part-Linear Multidimensional Projection*)

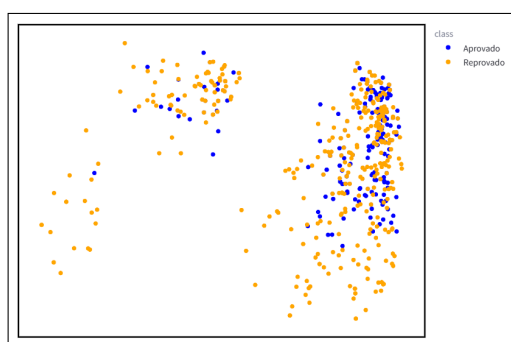


Figura 12 – Visualização de dados da disciplina GSI002 projetada com o modelo PLMP.

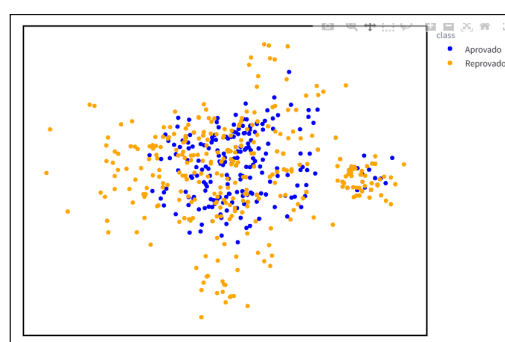


Figura 13 – Visualização de dados da disciplina GSI005 projetada com o modelo PLMP.

A Figura 12 apresenta a projeção PLMP de GSI002 com três aglomerados bem definidos, sendo o mais volumoso mais à direita e o menos volumoso mais à esquerda.

Apesar da definição dos grupos, percebe-se que há registros de ambas as classes em todos os três, o que significa que todos apresentam padrões de presenças muito parecidos.

Já a Figura 13 apresenta a projeção PLMP de GSI005 com um grande aglomerado que visivelmente possui subconjuntos menores, porém sem nenhum agrupamento específico para cada classe.

3.3.4 Modelo t-SNE (*t-Distributed Stochastic Neighbor Embedding*)

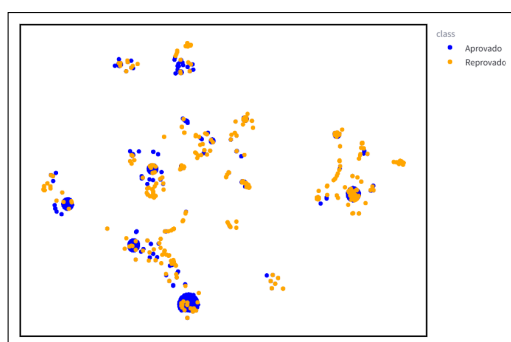


Figura 14 – Visualização de dados da disciplina GSI002 projetada com o modelo t-SNE.

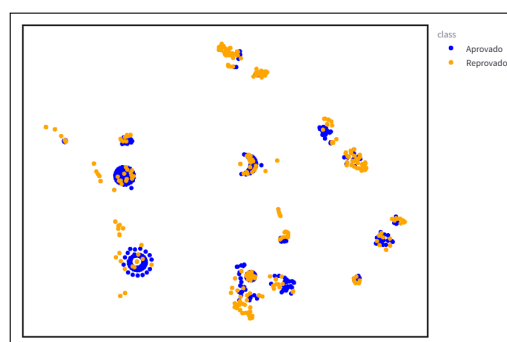


Figura 15 – Visualização de dados da disciplina GSI005 projetada com o modelo t-SNE.

A Figura 14 apresenta a projeção t-SNE de dados da disciplina GSI002, a qual visivelmente possui diversos pequenos aglomerados, cada um de uma combinação específica de presenças registradas e tópicos ministrados, portanto é possível verificar também quais grupos de presenças apresentam maioria aprovada ou reprovada, o que permite deduzir que cada agrupamento seja considerado padrão para uma das duas classes, porém não se pode verificar visualmente a dispersão da classe em si.

Já a Figura 15, que apresenta a projeção t-SNE de dados da disciplina GSI005, nota-se que as projeções obtiveram uma formação de aglomerados parecida, mas específica para a disciplina, portanto são resultados muito parecidos.

Nota-se a característica da base de que todos os alunos reprovados tiveram maior dispersão que em todas as projeções feitas, como há diversas situações onde o aluno pode ser reprovado, independente da presença em sala de aula, enquanto os alunos aprovados apresentaram menor dispersão e formação de agrupamentos, permitindo observar comportamentos parecidos entre os alunos aprovados, o que era esperado sobre essas bases de dados.

Porém, como os dados obtidos nessas técnicas de projeção podem agrupar mas não representar informações relevantes sobre o aglomerado de alunos, precisa-se de uma representação não apenas visual, mas também que permita observar a relação com os dados iniciais, para permitir observar que os dados obtidos coincidam corretamente.

4 Conclusão

Este trabalho analisou a viabilidade das projeções em representar os dados acadêmicos dos alunos nas projeções multidimensionais ISOMAP, LSP, PLMP e t-SNE, considerando a capacidade de representar os padrões das amostras de alunos usadas. Os modelos apresentaram uma boa representatividade nos resultados finais, o que possibilita observar os comportamentos sobre os grupos e quais as relações que eles apresentam com os dados.

Com a ferramenta de visualização, há certas diferenças entre os modelos:

- O modelo t-SNE apresentou agrupamentos extremamente compactos, o que é característico deste modelo, com distâncias quase equidistantes, o que causou sobreposição de coordenadas em uma visão global, mas maior precisão sobre os diferentes grupos formados. Com isso ela permite visão mais precisa sobre os padrões de presenças de alunos em certas matérias e turmas.
- O modelo ISOMAP possui uma representação mais linear de grupos, com uma boa representação de subgrupos, mas não permitiu representação tão precisa quanto a do t-SNE, porém permitiu verificar alguns dos grupos presentes, ainda com uma dispersão considerável.
- Os modelos LSP e PLMP não apresentaram tanta precisão na formação de aglomerados na maioria das disciplinas, o que dificulta a formação de grupos mais bem definidos e afeta a percepção de padrões de comportamento de discentes. apesar disso a visão global com foco na classificação dos alunos entre aprovado e reprovado tem certa relevância, mas apenas ao relacionar com os dados originais.

Os critérios de avaliação dos modelos incluíram preservação da estrutura global, clareza visual e a capacidade de revelar padrões nos dados projetados. Cada critério foi ponderado com base na sua relevância para a análise visual através da ferramenta desenvolvida, resultando em uma avaliação equilibrada.

Para a base utilizada, orienta-se primeiro o uso de técnicas com foco em observar os diferentes conjuntos de dados, o que permite maior precisão em agrupamentos específicos, primeiro com t-SNE e, caso seja preciso uma visão mais generalizada, é possível o uso de ISOMAP. Em alguns casos o uso de

Por fim, a ferramenta de visualização possibilita a análise dos dados da base inicial por meio da projeção do método escolhido, complementada por filtros específicos e, neste projeto, permite a avaliação dos diferentes modelos de projeção. Com aprimoramentos vol-

tados para gerar visões personalizadas para docentes ou discentes, e a integração de dados complementares como notas de avaliações e dados sobre repetentes, sendo viável realizar uma análise mais aprofundada tanto do desempenho individual quanto das disciplinas e o uso prático das análises para otimização de ementas, cargas horárias, entre outros, contribuindo significativamente para a melhor interpretação dos resultados educacionais.

Referências

- BALASUBRAMANIAN, M.; SCHWARTZ, E. L. The Isomap algorithm and topological stability. **Science**, v. 295, n. 5552, p. 27–34, jan. 2002. Disponível em: <https://doi.org/10.1126/science.295.5552.7a>. Citado na página 15.
- BORG, I.; GROENEN, P. J. F. **Modern multidimensional scaling: Theory and applications**. 2. ed. New York: Springer, 2005. 614 p. (Springer Series in Statistics). Disponível em: <https://doi.org/10.1007/0-387-28981-X>. Citado na página 12.
- BORGES, V. R. P.; ESTEVES, S.; ARAÚJO, P. de N.; OLIVEIRA, L. C. de; HOLANDA, M. Using principal component analysis to support students' performance prediction and data analysis. In: **Anais do XXIX Simpósio Brasileiro de Informática na Educação**. [s.n.], 2018. p. 1383–1392. Disponível em: <https://doi.org/10.5753/cbie.sbie.2018.1383>. Citado na página 17.
- FAYYAD, U.; PIATETSKY-SHAPIO, G.; SMYTH, P. From data mining to knowledge discovery in databases. **AI Magazine**, v. 17, n. 3, p. 37–54, set. 1996. Disponível em: <https://doi.org/10.1609/aimag.v17i3.1230>. Citado na página 12.
- FERNANDES, A. B.; NARCISO, R.; GUIMARÃES, C. D.; MATTOS, C. G. de; ERNANDES, I.; LIMA JUNIOR, I. N. de; BERTOLAZZI, J. C.; PARESCHI, S. C. S.; POLARI, S. de F. O.; SILVA, T. P. A. da. Inteligência artificial na avaliação de desempenho acadêmico: Desafios e oportunidades no ensino médio. **Revista Ibero-Americana de Humanidades, Ciências e Educação**, v. 10, n. 3, p. 180–196, mar. 2024. Disponível em: <https://doi.org/10.51891/rease.v10i3.13059>. Citado na página 17.
- FREITAS, C. M. dal S.; CHUBACHI, O. M.; LUZZARDI, P. R. G.; CAVA, R. A. Introdução à visualização de informações. **Revista de Informática Teórica e Aplicada**, v. 8, n. 2, p. 144–158, out. 2001. Disponível em: <http://hdl.handle.net/10183/19398>. Citado na página 11.
- GRÉGIO, A. R. A.; CARVALHO FILHO, B. P. de; MONTES, A.; SANTOS, R. Técnicas de visualização de dados aplicadas à segurança da informação. In: SANTIN, A.; NUNES, R. C.; DAHAB, R. (Ed.). **Minicuros do SBSEG 2009**. Porto Alegre, RS: SBC, 2009. cap. 5, p. 195–236. Citado 2 vezes nas páginas 9 e 12.
- MAATEN, L. van der; HINTON, G. Visualizing data using t-SNE. **Journal of Machine Learning Research**, v. 9, n. 86, p. 2579–2605, 2008. Disponível em: <http://jmlr.org/papers/v9/vandemaaten08a.html>. Citado na página 16.
- PAULOVICH, F. V.; NONATO, L. G.; MINGHIM, R.; LEVKOWITZ, H. Least square projection: A fast high-precision multidimensional projection technique and its application to document mapping. **IEEE Transactions on Visualization and Computer Graphics**, v. 14, n. 3, p. 564–575, maio 2008. Disponível em: <https://doi.org/10.1109/TVCG.2007.70443>. Citado na página 15.

PAULOVICH, F. V.; SILVA, C. T.; NONATO, L. G. Two-phase mapping for projecting massive data sets. **IEEE Transactions on Visualization and Computer Graphics**, v. 16, n. 6, p. 1281–1290, nov. 2010. Disponível em: <https://doi.org/10.1109/TVCG.2010.207>. Citado na página 16.

SILVA, E. V. da; MAGALHÃES NETTO, J. F. de; SOUZA, R. A. L. de. VLA dashboard: Um mecanismo para visualização do desempenho de estudantes de matemática no ensino médio. **Revista Novas Tecnologias na Educação**, UFRGS, v. 16, n. 2, p. 484–494, 2018. Disponível em: <https://doi.org/10.22456/1679-1916.89234>. Citado na página 17.

TENENBAUM, J. B.; SILVA, V. de; LANGFORD, J. C. A global geometric framework for nonlinear dimensionality reduction. **Science**, v. 290, n. 5500, p. 2319–2323, dez. 2000. Disponível em: <https://doi.org/10.1126/science.290.5500.2319>. Citado na página 14.

ZHOU, Y.; SHARPEE, T. O. Using global t-SNE to preserve inter-cluster data structure. **Neural Computation**, v. 34, n. 8, p. 1637–1651, jul. 2022. Disponível em: https://doi.org/10.1162/neco_a_01504. Citado 2 vezes nas páginas 5 e 14.