MINISTÉRIO DA EDUCAÇÃO
SECRETARIA DE EDUCAÇÃO PROFISSIONAL E TECNOLÓGICA
INSTITUTO FEDERAL DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA DO SUL DE MINAS GERAIS
CIÊNCIA DA COMPUTAÇÃO

Projeto de Trabalho de Conclusão de Curso

Avaliação da classificação de imagens hiperespectrais após a redução da dimensionalidade usando algoritmos de *Manifold Learning*

1.00.00.00-3 - Ciências Exatas e da Terra

1.03.00.00-7 - Ciência da computação

09 de Julho de 2021

Muzambinho – MG

INFORMAÇÕES GERAIS

Título do projeto: Avaliação da classificação de IHs após a redução da dimensionalidade usando algoritmos de Manifold Learning

Orientador(a) - Nome: Diego Saqui
E-mail: diego.saqui@ifsuldeminas.edu.br
Endereço no Lattes: http://lattes.cnpq.br/4408364907687419

Discente - Nome: Rafael Vicente da Silva
E-mail: 12151002569@muz.ifsuldeminas.edu.br
Endereço no Lattes: http://lattes.cnpq.br/7393602360594272

Membros do projeto:

Nome	Titulação máxima	Instituição Pertencente	Função	E-mail
Rafael Vicente da Silva	Bacharelando	IFSULDEMINAS - Campus Muzambinho	Orientado	12151002569@muz.ifsuldeminas.edu.br
Diego Saqui	Doutor	IFSULDEMINAS - Campus Muzambinho	Orientador	diego.saqui@muz.ifsuldeminas.edu.br

Local de Execução: IFSULDEMINAS – Campus Muzambinho.

Período de Execução:
Início: Fevereiro/2021
Término: Dezembro/2021

1. ANTECEDENTES, CARACTERIZAÇÃO DO PROBLEMA E JUSTIFICATIVA

O conceito de *Big Data* está relacionado à capacidade de processar e analisar grandes volumes de informação que permitam a extração de conhecimentos úteis para melhorar o processo de tomada de decisão (MARQUESONE, 2016), com isto os grandes volumes de dados podem ser caracterizados em relação a sua quantidade e também a quantidade de seus atributos (dimensões) (JACOBS, 2009; LAZER et al., 2014). Desta forma, surgem técnicas envolvendo o desenvolvimento e aplicação de métodos de reconhecimento de padrões e mineração de dados (DISNER, 2015). Exemplos de sistemas que operam com dados de alta dimensão e/ou *Big Data* incluem áreas como análise de dados geoespaciais, bioinformática, organização e recuperação de imagens baseada em conteúdo, bases de dados distribuídas na internet, redes de sensores e imagens hiperespectrais (IHs).

Em especial, as IHs ilustram a composição química por meio de imagens feitas a partir de informações espectrais coletadas por um espectrômetro (Figura 1), cujo sensor hiperespectral capta milhares ou centenas de milhares de espectros, ao invés de um único espectro (GHAMISI et al., 2017). Pode-se mostrar que classes espectralmente muito semelhantes, isto é, classes que compartilham de vetores de médias, muito próximos entre si, podem ser separadas com um grau de acurácia em espaços de dimensão suficientemente alta (FUKUNAGA, 1990).

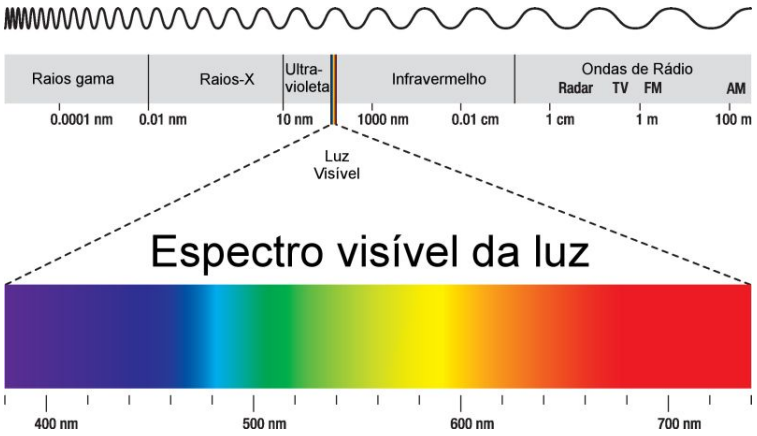

Figura 1: Comportamento do comprimento de onda de espectro.

Em razão do avanço da computação e dos sistemas sensores, surgiram novas possibilidades de manipulação no domínio espectral por meio do Sensoriamento Remoto (SR), ao qual Florenzano (2007) define SR como sendo a tecnologia que possibilita obter imagens - e outros tipos de dados - da superfície terrestre, através da captação e do registro da energia refletida ou emitida pela superfície. O termo sensoriamento está relacionado à obtenção dos dados por meio de sensores situados em plataformas terrestres, aéreas e orbitais,

e o termo remoto, que significa distante, é utilizado pois a obtenção é feita sem o contato físico do sensor ao objeto de estudo, como ilustrado na Figura 2.

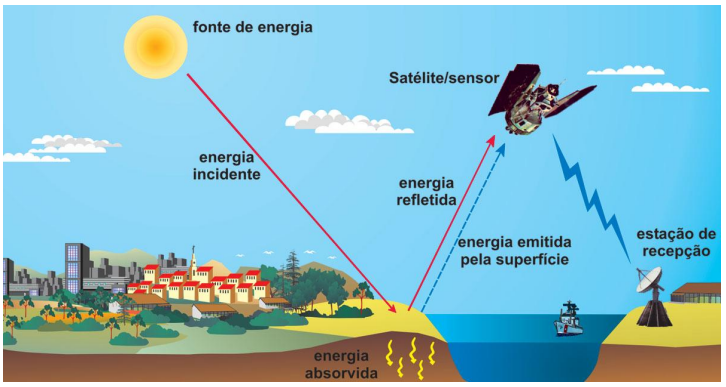

Figura 2: Obtenção de imagens por SR. Extraída de (FLORENZANO, 2007)

Utilizando sensores de alta resolução espectral, que proporcionam para cada pixel (elemento de resolução espacial), medidas radiométricas em bandas estreitas e contínuas, pode-se obter uma grande quantidade de informações espectrais em seu domínio. Essas informações têm um nível de resolução mais próximo daquele verificado em espectrorradiômetros de campo ou de laboratório, facilitando o uso de abordagens mais específicas, que permitam quantificar alvos com maior nível de detalhamento espectral, compondo assim as imagens hiperespectrais (CLARK, 1999).

IHs geralmente são vistas como cubos hiperespectrais, onde as imagens de banda simples estão empilhadas de modo que a terceira dimensão do cubo é incrementada pelos comprimentos de onda amostrados. Como demonstra a Figura 3.

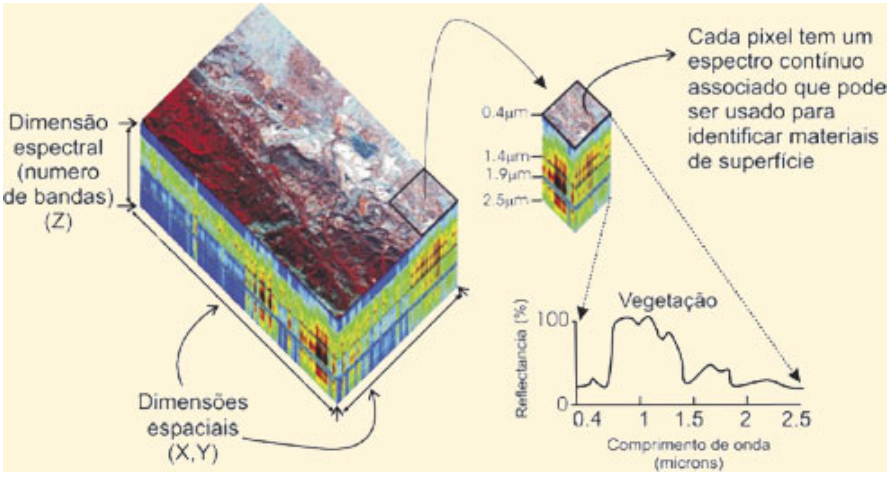

Figura 3: Ilustração do comportamento de IHs. Extraída de (MUNDOGEO, 2021)

Uma IH que é coletada em *d* bandas espectrais pode ser pensada como uma nuvem de pontos (dados pelos pixels na imagem) em um espaço *d*-dimensional. As *d* bandas espectrais na imagem formam *d* eixos de coordenadas no hiperespaço. Cada pixel é representado como um vetor *d*-dimensional tal que o valor da i-ésima coordenada é o valor estimado da reflectância terrestre (ou da radiância do alcance do sensor) medida no i-ésimo comprimento de onda. Esse vetor é a representação geométrica do pixel no espaço espectral.

A complexidade e o grande volume de dados inerentes, exigem *software* adequado para a sua análise (ZHANG *et al*., 2000) e a utilização de algoritmos de classificação apropriados. Entre eles pode-se citar os sensores *Airbone Visible Infrared Imager Spectroradiometer* ("AVIRIS"), com 224 bandas espectrais e o HYPERION a borda do satélite Hyperspectral Imager *(Earth Observing 1)*, com 220 bandas espectrais.

Jimenez et al., (1998) aponta que o impacto do problema da complexidade da dimensionalidade dos dados varia de um campo para outro. Para a otimização combinatória de muitas dimensões é visto como um crescimento exponencial do esforço computacional, já no campo estatístico se manifesta como um problema de parâmetro ou densidade estimativa, devido à escassez de dados.

Os dados/registros computacionais que possuem uma quantidade elevada de atributos (dimensões), dizemos que são dados superdimensionados. Dados superdimensionados oferecem um poder discriminante mais elevado do que dados com baixa dimensionalidade, contudo a análise destes pode evidenciar um problema chamado de maldição da dimensionalidade ou fenômeno de Hughes (HUGHES, 1968).

A maldição da dimensionalidade é um fator desafiador na modelagem matemática visto que, para um hiperplano cartesiano com *d* dimensões de entrada onde cada dimensão de entrada é particionada em *s* células, o número total de células seria s^d (BELLMAN, 1961). Como consequência disso, a criação de modelos destes dados necessita considerar espaços de busca inerentemente esparsos (LAROSE, 2006). Desta forma, os cientistas têm se deparado com a necessidade de encontrar estruturas significativas ocultas de baixa dimensão, dentro de dados de alta dimensão, sendo tal técnica denominada de redução de dimensionalidade dos dados (RDD), (PEARSON, 1901; HOTELLING, 1933; JOLLIFE, 2003; ROWEIS et al., 2003; DONOHO et al., 2003).

O efeito negativo desta escassez resulta de alguma geometria, estatística e propriedades assintóticas do espaço de recursos de alta dimensão, e essas características exibem um comportamento surpreendente para dados em dimensões superiores. Os autores

Kendall (1961) e Scott (1992) descrevem que a concentração de dados no hiperespaço à medida que a dimensionalidade aumenta, tendem a ficarem isolados em determinados pontos do hiperespaço. Desta maneira conclui-se que o hiperespaço de certa forma acaba sendo um "grande vazio" e os dados concentrando-se nas bordas, evidencia-se o fato da simplificação do problema usando da redução de dimensionalidade e ainda a correlação dos dados de alta dimensão aproxima-se muito para dados de baixa dimensão.

A partir disso, devido ao grande volume de dados (bandas espectrais de IHs), a extração destas informações não é uma tarefa trivial, onde são necessários o uso de teorias e ferramentas para o auxílio na extração e análise de informações úteis, facilitando assim a classificação (BORGES et al., 2006). Porém, para tratar de dados com alta dimensionalidade, podemos utilizar estratégias baseadas em *Manifold Learning. Manifold Learning* partem essencialmente da distância (correlação) que os dados estão dispostos entre si no espaço, o grau de afinidade nos dados, e demais métricas irão reduzir a dimensionalidade da base de dados e simultaneamente preservar a relação dos mesmos (CAYTON, 2005).

Considerando o contexto anterior, este trabalho pretende fazer o uso de algoritmo de *Manifold Learning,* para a redução da dimensionalidade de modo a obter uma base de dados simplificada, facilitando desta maneira a extração de informações a respeito da mesma, possibilitando que os classificadores obtenham resultados satisfatórios. O propósito é otimizar o tempo de classificação de uma IHs e tendo uma classificação assertiva dos dados.

2. REFERENCIAL TEÓRICO

Este capítulo apresenta os principais conceitos a serem utilizados neste projeto e trabalhos relacionados. Inicialmente, é destacado o comportamento dos dados no hiperespaço. Na sequência têm-se os algoritmos a serem empregados, sendo dois redutores, análise de componentes principais, do inglês *Principal Components Analytics* (PCA), é uma técnica multivariada de modelagem da estrutura de covariância e Incorporação de vizinhos estocásticos com distribuição t, do inglês *t-Distributed Stochastic Neighbor Embedding* (t-SNE), na qual converte as distâncias entre os pontos no espaço multidimensional em probabilidades que representam as similaridades, e dois classificadores, K-ésimo vizinho mais próximo, do inglês *K-Nearest Neighbors* (KNN), cuja a classificação é feita por meio da busca dos k vizinhos, utilizando uma medida de distância nesta procura e suporte de máquina vetorial, do inglês *Support Vector Machine* (SVM), têm como objetivo a determinação de limites de decisão que produzam uma separação ótima entre classes.

2.1. Demonstração matemática sobre a dimensionalidade dos dados

Nesta seção, ilustramos alguns fatos incomuns ou inesperadas características do hiperespaço (espaços para mais de três dimensões), incluindo uma prova e discussão. Tais ilustrações pretendem mostrar que o espaço dimensional superior é bastante diferente do espaço 3-D com o qual nós somos familiares.

Conforme a dimensionalidade aumenta:

A. O volume de um hipercubo concentrado nos cantos (SCOTT, 1992)

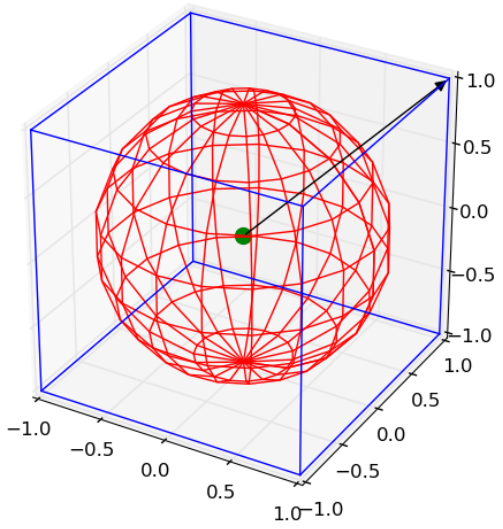

Figura 4: Comportamento dos dados no hiperespaço.

Fonte: Do autor

Foi demonstrado (KENDALL, 1961) que o volume da hiperesfera de raio *r* e dimensão *d* é dado por:

$$V_s(r) = \text{volume da hiperesfera} = \frac{2r^d}{d}\,\frac{\pi^{d/2}}{\Gamma\left(\frac{d}{2}\right)} \qquad (1)$$

e que o volume de um hipercubo em [*-r, r*]^d é dado por

$$V_c(r) = \text{volume do hipercubo} = (2\mathrm{r})^d. \qquad (2)$$

A fração do volume de uma hiperesfera inscrita em um hipercubo é

$$f_{d1} = \frac{V_s(r)}{V_c(r)} = \frac{\pi^{d/2}}{d2^{d-1}\Gamma(d/2)} \qquad (3)$$

onde *d* é o número de dimensões. Vemos na Fig. 5 como f_{d1} diminui à medida que a dimensionalidade aumenta.

Observe que $\lim_{d\to\infty} f_{d1} = 0$, o que implica que o volume do hipercubo está cada vez mais concentrado nos cantos conforme *d* aumenta, conforme demonstrado na Fig. 5.

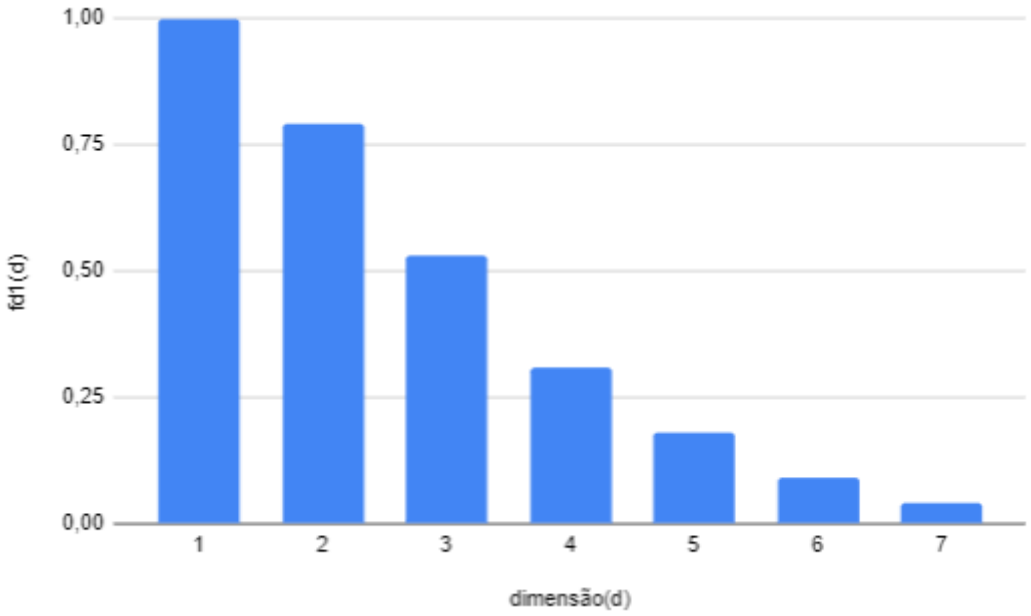

Figura 5. Volume fracionário de uma hiperesfera inscrita em um hipercubo como um função da dimensionalidade.

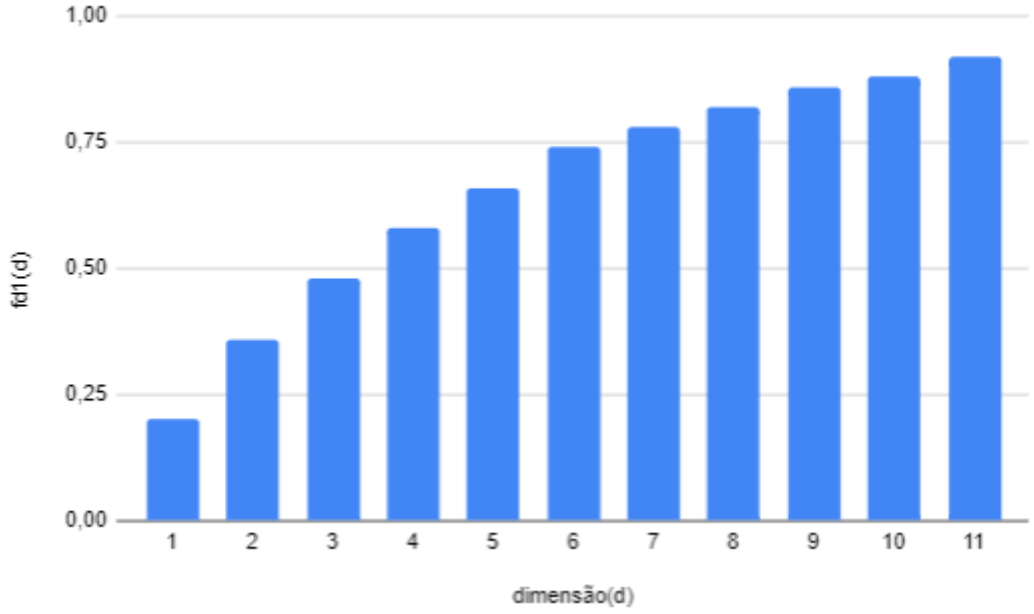

Figura 6. Volume de uma hiperesfera contida na casca externa como uma função de dimensionalidade para $\xi = r/5$

As características mencionadas anteriormente têm duas consequências importantes para dados de alta dimensão que aparecem imediatamente. O primeiro é que o espaço de alta dimensão é quase vazio, o que implica que os dados multivariados geralmente possuem uma estrutura dimensional inferior. Como consequência, dados de alta dimensão podem ser projetados para uma dimensão inferior ao subespaço sem perder informações significativas, em termos de separabilidade entre as diferentes classes estatísticas. A segunda consequência do exposto é que os dados normalmente distribuídos terão tendência a se concentrar nas caudas; similarmente, dados uniformemente distribuídos terão maior probabilidade de serem coletados nos cantos, tornando a estimativa da densidade mais difícil. As vizinhanças locais estão quase certamente vazias, exigindo que a largura de banda de estimativa seja grande e produzindo o efeito de perder estimativa de densidade detalhada.

Suporte para esta tendência pode ser encontrado nas estatísticas, o comportamento de multivariadas normalmente são uniformemente distribuídas para dados em alta dimensionalidade. Espera-se que, à medida que a dimensionalidade aumenta, os dados se concentrem em uma casca externa.

2.2. Algoritmos empregados

2.2.1. Redutor: Análise de Componentes Principais

A análise de componentes principais, do inglês Principal Components Analytics (PCA), é uma técnica multivariada de modelagem da estrutura de covariância. A técnica foi inicialmente descrita por Pearson (1901) e uma descrição de métodos computacionais práticos veio muito mais tarde com Hotelling (1933, 1936) que usou com o propósito determinado de analisar as estruturas de correlação. A PCA é uma técnica estatística de análise multivariada que transforma linearmente um conjunto original de variáveis, inicialmente correlacionadas entre si, num conjunto substancialmente menor de variáveis não correlacionadas que contém a maior parte da informação do conjunto original.

PCA é a técnica mais conhecida e está associada à ideia de redução de massa de dados, com menor perda possível da informação, contudo é importante ter uma visão conjunta de todas ou quase todas as técnicas da estatística multivariada para resolver a maioria dos problemas práticos, também é associada à ideia de redução de massa de dados, com menor perda possível da informação. Procura-se redistribuir a

variação observada nos eixos originais de forma a se obter um conjunto de eixos ortogonais não correlacionados (MANLY, 1986; HONGYU, 2015).

O objetivo principal da análise de componentes principais é o de explicar a estrutura da variância e covariância de um vetor aleatório, composto de p-variáveis aleatórias, por meio de combinações lineares das variáveis originais. Essas combinações lineares são chamadas de componentes principais, que são as dimensões, e são não correlacionadas entre si (SANDANIELO, 2008).

2.2.2. Redutor: Incorporação de vizinhos estocásticos com distribuição t

O algoritmo t-SNE, da classe de Manifold Learning fornece um método eficaz para visualizar um conjunto complexo de dados. Ele descobre com sucesso estruturas ocultas nos dados, expondo clusters naturais e suavizando variações não-lineares ao longo das dimensões, reduzindo para duas ou três dimensões (VAN DER MAATEN, 2008). Muitos conjuntos de dados do mundo real têm uma baixa dimensionalidade intrínseca, apesar de estarem inseridos em um espaço de alta dimensão. Esse espaço de baixa dimensão está embutido no espaço de alta dimensão de uma maneira complexa e não linear. Escondido nos dados, esta estrutura só pode ser recuperada através de métodos matemáticos específicos (ROSSANT, 2015).

O método, em síntese, converte as distâncias entre os pontos no espaço multidimensional em probabilidades que representam as similaridades. No espaço dimensional reduzido, as distâncias são também calculadas, sendo posteriormente ajustadas conforme o cálculo do gradiente, que representa a similaridade posicional dos pontos em relação a ambos os espaços dimensionais.

2.2.3. Classificador: K - *Nearest Neighbors*

O KNN é um dos algoritmos de classificação mais utilizados na área de aprendizagem de máquina (DINIZ et., 2013). É baseado na procura dos k vizinhos mais próximos do padrão de teste. A busca pela vizinhança é feita utilizando uma medida de distância nessa procura.

O KNN foi proposto por Fukunaga e Narendra (1975), este é um classificador onde o aprendizado de um novo “objeto” é feito com base nos exemplos de treinamento (aprendizagem supervisionada). Onde pode ser expresso por:

$$d(X_i, Yi) = \sqrt[r]{\sum_{i=1}^{n} |X_i - Yi|^r}$$

Onde *d* representa a distância, e *Xi* e *Yi* representam as instâncias, *n* é o número de atributos e *r* sendo a dimensão pertencente, podendo ocorrer certas alterações dependendo da métrica utilizada, Euclidiana, Manhattan ou então Minkowski.

2.2.4. Classificador: Support Vector Machine

Support Vector Machines (SVM) é um algoritmo supervisionado baseado na teoria do aprendizado estatístico (teoria Vapnik-Chervonenkis) projetado para tarefas de classificação, que têm como objetivo a determinação de limites de decisão que produzam uma separação ótima entre classes por meio da minimização dos erros (VAPNIK et al., 1995).

O SVM consiste em uma técnica computacional de aprendizado para problemas de reconhecimento de padrão. Introduzida por meio da teoria estatística de aprendizagem por Vapnik et al., (1995), essa classificação é baseada no princípio de separação ótima entre classes, tal que se as classes são separáveis, a solução é escolhida de forma a separar o máximo as classes.

O SVM busca um hiperplano ótimo como uma função de decisão em um espaço de características que pode ter muitas dimensões (BOSER et al., 1992; CRISTIANINI; SHAWE-TAYLOR, 2000). Para essa otimização, o SVM introduz uma minimização do risco estrutural, do inglês structural risk minimization (SRM), considerando o melhor separador, aquele que minimiza o erro de generalização e tentando evitar problemas de *overfitting* (GUO et al., 2019).

2.3. Trabalhos relacionados

2.3.1. Um novo método Wrapper multiobjetivo para seleção de bandas de Imagens Hiperespectrais, por Saqui (2020)

Nesta pesquisa foi elaborado um método de seleção de bandas multiobjetivo chamado Wrapper Multiobjective Evolutionary Band Selecion (WMoEBS) composto por estratégias que foram testadas experimentalmente. O WMoEBS é baseado na estratégia Wrapper incorporando o classificador Support Vector Machine (SVM), que utiliza informação espacial e espectral, e realiza uma seleção inicial para diminuir as bandas correlacionadas, consistindo num algoritmo multiobjetivo para lidar com

resultados da classificação e quantidade de bandas simultaneamente e um tomador de decisão para retornar uma única solução final.

2.3.2. Extração de atributos em imagens de sensoriamento remoto utilizando Independent Component Analysis e combinação de métodos lineares, por Levada (2006)

O presente trabalho de Levada, apresenta uma metodologia para melhorar o desempenho da classificação criando modelos para fusão de atributos que combinam métodos estatísticos de segunda ordem com métodos de ordens superiores, superando limitações existente nas abordagens tradicionais, como problemas de mal-condicionamento, o que pode provocar instabilidade na estimação dos componentes independentes, além de eventuais amplificações de ruídos. O esquema resultante é utilizado para combinar atributos obtidos através de diversos métodos num único vetor de padrões em duas abordagens: Fusão Concatenada e Fusão Hierárquica

2.3.3. Redução de dimensionalidade em imagens hiperespectrais usando Codificadores automáticos, por Kakarla et al., (2020)

Este artigo apresenta uma análise não linear para redução de dimensionalidade usando codificadores automáticos. O desempenho do modelo proposto é comparado com outros métodos popularmente usados como PCA e kernel PCA (KPCA) usando os classificadores KNN, KNN ponderado em conjuntos de dados de benchmark obtidos da Repositório Computational de dados Intelligence Group (CIG). Experimentalmente, foi provado que a técnica proposta usando auto-codificadores supera as técnicas de redução de dimensionalidade existentes PCA e KPCA.

2.3.4. Demais trabalhos

O emprego de IHs em estudos florestais vem crescendo à medida que aumenta a exigência de um detalhamento maior da estrutura das florestas, de modo a ser cada vez mais eficiente, sendo possível a discriminação e reconhecimento de características dos vegetais, guiando a uma análise mais precisa da composição e condições sanitárias da floresta (PETEAN, 2015).

Outro uso para IHs está voltado ao ambiente urbano, possibilitando a distinção entre os mais diversos componentes da paisagem urbana de maneira confiável (PETEAN, 2015). De acordo com Herold et al. (2003), os ambientes urbanos representam uma das mais desafiadoras áreas de análise para o sensoriamento remoto, pois sua diversidade espectral excede àquela encontrada nos ambientes naturais.

E, além disso, têm-se processos aos quais usam da espectroscopia de IHs e a quimiometria para a quantificação e classificação no ramo da agroindústria, como: identificação de fontes de licopeno e carotenóides, como o betacaroteno (BARANSKA et al., 2006), classificação de azeite extravirgem (SINELLI et al., 2010), determinação de parâmetros de qualidade em produtos lácteos (RŮŽIČKOVÁ; SUSTOVÁ, 2006), caracterização de azeitona de mesa (CASALE et al., 2010), determinação de açúcar em uva (JARÉN et al., 2001), quantificação do teor de proteína em produtos de leite em pó (INGLE et al., 2016), controle de qualidade de extratos de frutos silvestres durante o armazenamento (GEORGIEVA et al., 2014).

3. OBJETIVOS

3.1. Objetivo Geral

O presente trabalho tem por objetivo avaliar a redução da dimensionalidade dos dados em imagens hiperespectrais (IHs) a partir do uso de algoritmos de *manifold learning*, explorando a maldição da dimensionalidade, e com isto analisar a classificação de IHs, verificando o comportamento dos algoritmos empregados, a fim de investigar o que melhor se aplicam à base de dados específica de estudo, *Indian Pines,* analisando a precisão do resultado juntamente ao tempo empregado na classificação, de modo a investigar a eficiência.

3.2. Objetivos Específicos

- Realizar comparações entre os algoritmos de manifolds learning, a fim de identificar o qual melhor aplica-se a base de *Indian Pines*.
- Aplicar os redutores PCA (Análise de Componentes Principais) e t-SNE (t-distributed Stochastic Neighbor Embedding), a fim de identificar o qual melhor aplica-se na base de *Indian Pines*.
- Aplicar os classificadores KNN (K-Nearest Neighbors) e SVM (Support Vector Machines), a fim de identificar qual melhor aplica-se à base de *Indian Pines* .

- Realizar comparações entre os classificadores, de modo a apurar qual obterá melhor resultado classificando a IHs de *Indian Pines*.
- Gerar uma arquitetura que represente o procedimento completo utilizado neste estudo.

4. METODOLOGIA E ESTRATÉGIA DE AÇÃO

A metodologia foi conduzida conforme as etapas estabelecidas por Fayyad et al. (1996): Pré-processamento, Transformação, Mineração de dados, Avaliação e Interpretação (Figura 6).

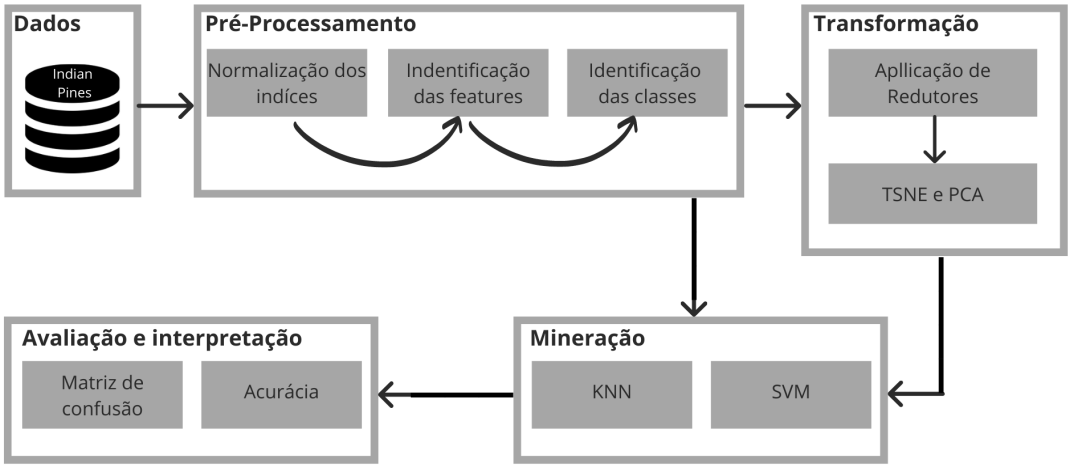

Figura 7. Fluxograma descrevendo as etapas de Análise e Descoberta de Conhecimento (KDD) para as bases de dados Indian Pines.

Fonte: Do autor.

4.1. Dados

A área sob estudo foi coletada em Junho de 1992 pelo sensor AVIRIS na região noroeste do estado americano de Indiana, a cena identificada como Indian Pines é composta de 145 × 145 pixels e 224 bandas de refletância com comprimento de onda no intervalo de 0,4 a 2,5 µm, sendo que a base de dados utilizada é a imagem tratada em *Comma-separated values* (CSV) com 200 bandas espectrais, ao qual cada amostra de dados para os classificadores é um pixel da imagem Indian Pines. Uma visualização em tons de cinza é dada na Figura 8.a. Uma visualização em falsa composição *Red Green Blue* (RGB) é dada na Figura 8.b. Uma visualização da rotulação da região é dada na Figura 8.c. A identificação dos rótulos da IH é dada na Tabela 1.

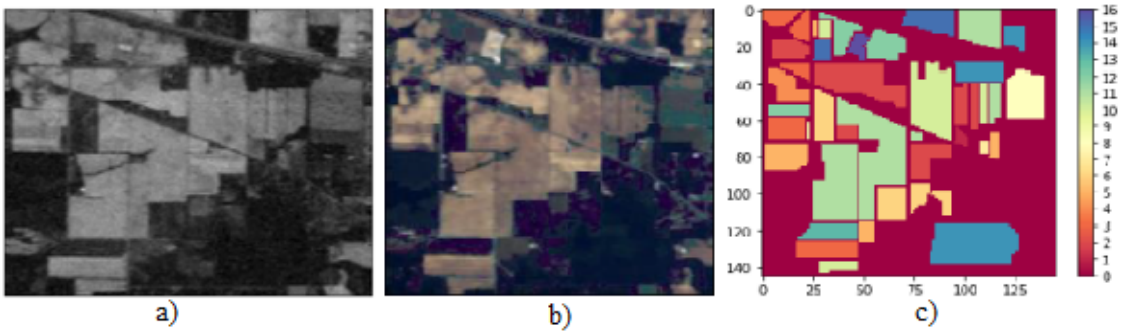

Figura 8: a) Indian Pines - imagem em tons de cinza. Extraída de (BAUMGARDNER; BIEHL; LANDGREBE, 2015) b) Indian Pines - falsa composição RGB. Extraída de (SAQUI, 2018) c) Indian Pines - mapa de rótulos

Rótulo	Descrição
1	Alfafa
2	Milho - primeira fase
3	Milho - segunda fase
4	Milho - terceira fase
5	Grama - pastagem
6	Grama - árvores
7	Grama - pastagem cortada
8	Feno
9	Aveia
10	Soja - primeira fase
11	Soja - segunda fase
12	Soja - terceira fase
13	Trigo
14	Bosques
15	Construções - Grama - Árvores - Ruas
16	Rochas - Estruturas Férreas - Edifícios

Tabela 1: Indian Pines - tabela de classes

4.2. Pré-Processamento

O pré-processamento foi a parte responsável pelo tratamento e identificação dos dados a serem analisados na imagem (pixels), de modo a tratar as entradas para serem os parâmetros nos redutores e classificadores. E para a produção do projeto será utilizado a ferramenta *Google Colab*, voltada para o desenvolvimento dos algoritmos feitos na linguagem Python, na versão 3.8, utilizando as bibliotecas: Pandas para a análise dos dados, Numpy para cálculo em vetores multidimensionais, Matplotlib e Seaborn para representação gráfica dos dados, e também a Sklearn para o uso dos redutores e classificadores.

4.3. Transformação

A transformação pautou-se em reduzir a dimensionalidade da IH de *Indian Pines* com o PCA, originalmente a imagem com 200 bandas espectrais (dimensões) reduzindo para apenas 10 componentes (dimensões).

E a redução utilizando o t-SNE foi feita reduzindo originalmente a IH com 200 dimensões para apenas duas dimensões, tendo como parâmetros, 10 de perplexidade em 1200 iterações, reduzindo a base em 99% dos atributos originais.

4.4. Mineração

A mineração da imagem, dividira-se em classificar sem antes reduzir a dimensionalidade e, depois classificar a IH devidamente com a dimensionalidade reduzida, para assim realizar comparativos com os dados dimensionalmente reduzidos. Para isto fez-se uso do KNN com métrica euclidiana, analisando os pixels vizinhos de um intervalo de 100 à 1000, e obtendo a acurácia pela média da iteração, e já o SVM operou com uma taxa de regularização de 3500, utilizando um kernel radial, com coeficiente de kernel escalar, e ocupando 16 gigabytes de cache para a execução .

4.5. Avaliação e Interpretação

A avaliação e interpretação foi realizada através da geração de matriz de confusão para compreendimento da classificação e verificando a acurácia obtida, por meio das funções de Metrics da biblioteca Sklearn, e ainda utilizando a função de relatório de classificação, será possível também verificar quantos pixels fora classificado para cada uma das 16 classes, e quantos pertencem às mesmas. Deste modo será possível analisar o comportamento do

classificador para cada resultado de classe e pixel, verificando assim para qual está mais assertivo.

5. RESULTADOS E IMPACTOS ESPERADOS

Um dos aspectos em que ainda há um caminho longo para evolução na aprendizagem de máquina é em relação ao tratamento de dados com alta dimensionalidade, nesse sentido, o trabalho apresentou os redutores de dimensionalidade, t-SNE e PCA, com os classificadores KNN e SVM aplicados a imagem hiperespectral Indian Pines.

5.1. Resultados Alcançados

Abaixo apresenta-se a tabela da média da acurácia e tempo em segundos da classificação de Indian Pines e ainda acurácia e tempo esperado é obtido pela média dos trabalhos relacionados do referencial teórico, utilizando os classificadores KNN e SVM, com e sem os redutores de dimensionalidade, PCA e t-SNE.

	Obtido	**Esperado**	**Obtido**	**Esperado**
Algoritmo	**Acurácia da Classificação**		**Tempo(s)**	
KNN	69,61%	> 70,00%	3,12	≤
SVM	83,11%	> 85,00%	44,63	≤
PCA + KNN	57,30%	> 65,00%	1,34	≤
PCA + SVM	62,50%	> 75,00%	128,7	≤
t-SNE + KNN	59,21%	> 70,00%	0,91	≤
t-SNE + SVM	70,81%	> 90,00%	6,08	≤

Tabela 2: Média da acurácia e tempo da classificação de Indian Pines

Evidentemente o classificador SVM obteve um resultado proeminente mais satisfatório do que em relação aos demais, em questão de o mesmo têm uma maior gama de parâmetros para a classificação, necessitando assim de otimização dos parâmetros para o caso, o que já não ocorre ao KNN, pois o mesmo leva em conta apenas a distância de seus vizinhos para a classificação, que no estudo fora estimado da média seus seus primeiros 1000 vizinhos (ao qual representa pouco menos de 5% da amostra total), tendo carência na precisão nos resultados colhidos em questão da redução de dimensionalidade. Contudo, ao analisarmos o

tempo de execução da classificação do SVM com o redutor t-SNE, consegue um tempo de praticamente 7 vezes menor do que o mesmo sem o redutor da classe de manifold learning , o com o classificador KNN, e quando aplicado o t-SNE com KNN, consegue-se um melhor resultado em se tratando de precisão e tempo, obtendo 74,62% de acurácia com um tempo médio de 0,2 segundos por iteração.

Portanto, a partir dos resultados obtidos evidencia o fato do uso de RD para o caso de Indian Pines em específico, de modo a obter uma classificação mais rápida e medianamente precisa, e para a otimização da classificação é explícito que o uso do algoritmo de manifold learning, t-SNE melhora a acurácia e ainda consegue ser mais rápido na classificação.

6. CRONOGRAMA

Atividades	**2021**										
	Fev	**Mar**	**Abr**	**Mai**	**Jun**	**Jul**	**Ago**	**Set**	**Out**	**Nov**	**Dez**
Obtenção da IH de *Indian Pines*	x	x	x								
Pré-Processamento da IH, normalizando índices, identificando features e classes		x	x	x	x						
Mineração dos dados ainda não reduzidos a dimensionalidade com KNN				x	x	x	x				
Mineração dos dados ainda não reduzidos a dimensionalidade com SVM					x	x	x	x			
Transformação da IH aplicando redutor PCA						x	x	x			
Transformação da IH aplicando redutor t-SNE							x	x	x		
Mineração com KNN dos dados com dimensionalidade reduzida							x	x	x	x	
Mineração com SVM dos dados com dimensionalidade reduzida								x	x	x	
Avaliação e Interpretação dos resultados por meio da				x	x	x	x	x	x	x	x

análise da acurácia e geração da matriz de confusão											

8. REFERÊNCIAS BIBLIOGRÁFICAS

BARANSKA, Malgorzata; SCHÜTZE, W.; SCHULZ, Hartwig. **Determination of lycopene and β-carotene content in tomato fruits and related products: comparison of FT-Raman, ATR-IR, and NIR spectroscopy.** Analytical Chemistry, v. 78, n. 24, p. 8456-8461, 2006.

BAUMGARDNER, Marion F.; BIEHL, Larry L.; LANDGREBE, David A. **220 band aviris hyperspectral image data set: June 12, 1992 indian pine test site 3**. Purdue University Research Repository, v. 10, p. R7RX991C, 2015.

BELLMAN, Richard E. **Adaptive control processes: a guided tour.** Princeton university press, 2015.

BORGES, HELYANE BRONOSKI; NIEVOLA, J. C. Redução de Dimensionalidade em Bases de Dados de Expressão Gênica. 2006. Tese de Doutorado. Dissertação de Mestrado, PPGIa-PUCPR.

BOSER, Bernhard E.; GUYON, Isabelle M.; VAPNIK, Vladimir N. **A training algorithm for optimal margin classifiers.** In: Proceedings of the fifth annual workshop on Computational learning theory. 1992. p. 144-152.

CASALE, Monica et al. **Characterisation of table olive cultivar by NIR spectroscopy.** Food chemistry, v. 122, n. 4, p. 1261-1265, 2010.

CAYTON, Lawrence. **Algorithms for manifold learning.** Univ. of California at San Diego Tech. Rep, v. 12, n. 1-17, p. 1, 2005.

CLARK, Roger N. et al. **Spectroscopy of rocks and minerals, and principles of spectroscopy.** Manual of remote sensing, v. 3, n. 3-58, p. 2-2, 1999.

CRISTIANINI, Nello; SHAWE-TAYLOR, John. **Support Vector Machines and other kernel-based learning methods**. Cambridge, 2004.

DINIZ, Fábio Abrantes et al. **RedFace: um sistema de reconhecimento facial baseado em técnicas de análise de componentes principais e autofaces.** Revista Brasileira de Computação Aplicada, v. 5, n. 1, p. 42-54, 2013.

DISNER, Daniel da Silva. **Mineração de dados para obtenção de conhecimento em Big Data.** 2015.

DONOHO, David L.; GRIMES, Carrie. **Hessian eigenmaps: Locally linear embedding techniques for high-dimensional data. Proceedings of the National Academy of Sciences**, v. 100, n. 10, p. 5591-5596, 2003.

FAYYAD, Usama; PIATETSKY-SHAPIRO, Gregory; SMYTH, Padhraic. **From data mining to knowledge discovery in databases.** AI magazine, v. 17, n. 3, p. 37-37, 1996.

FLORENZANO, Teresa Gallotti. **Iniciação em sensoriamento remoto.** Oficina de textos, 2007.

FUKUNAGA, Keinosuke; NARENDRA, Patrenahalli M.. . **A branch and bound algorithm for computing k-nearest neighbors.** IEEE transactions on computers, v. 100, n. 7, p. 750-753, 1975.

FUKUNAGA, Keinosuke. **Introduction to statistical pattern recognition.** Elsevier, 2013.

GEORGIEVA, Mariya et al. **Application of NIR spectroscopy and chemometrics in quality control of wild berry fruit extracts during storage.** Hrvatski časopis za prehrambenu tehnologiju, biotehnologiju i nutricionizam, v. 8, n. 3-4, p. 67-73, 2013.

GHAMISI, Pedram et al. **Advances in hyperspectral image and signal processing: A comprehensive overview of the state of the art**. IEEE Geoscience and Remote Sensing Magazine, v. 5, n. 4, p. 37-78, 2017.

GUO, Yanhui et al. **Hyperspectral image classification with SVM and guided filter.** EURASIP Journal on Wireless Communications and Networking, v. 2019, n. 1, p. 1-9, 2019.

HEROLD, Martin et al. **The spectral dimension in urban land cover mapping from high-resolution optical remote sensing data**. In: Proceedings of the 3rd Symposium on remote Sensing of Urban Areas. 2002. p. 2002.

HONGYU, Kuang. **Comparação do GGE-biplot ponderado e AMMI-ponderado com outros modelos de interação genótipo × ambiente.** 2015. Tese de Doutorado. Tese (Doutorado em Estatística e Experimentação Agronômica). Piracicaba: USP. 155p.

HOTELLING, Harold. **Analysis of a complex of statistical variables into principal components.** Journal of educational psychology, v. 24, n. 6, p. 417, 1933.

HOTELLING, Harold. **Simplifield calculation of principal components**. Psychometrika, Williamsburg, v.1, p.27-35, 1936.

HUGHES, Gordon. **On the mean accuracy of statistical pattern recognizers.** IEEE transactions on information theory, v. 14, n. 1, p. 55-63, 1968.

INGLE, Prashant D. et al. **Determination of protein content by NIR spectroscopy in protein powder mix products.** Journal of AOAC International, v. 99, n. 2, p. 360-363, 2016.

JACOBS, Adam. **The pathologies of big data**. Communications of the ACM, v. 52, n. 8, p. 36-44, 2009.

JARÉN, C. et al. **Sugar determination in grapes using NIR technology.** International Journal of Infrared and Millimeter Waves, v. 22, n. 10, p. 1521-1530, 2001.

JIMENEZ, Luis O.; LANDGREBE, David A. **Supervised classification in high-dimensional space: geometrical, statistical, and asymptotical properties of multivariate data.** IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), v. 28, n. 1, p. 39-54, 1998.

JOLLIFFE, Ian T. **Principal component analysis.** Technometrics, v. 45, n. 3, p. 276, 2003.

KAKARLA, Syam et al. **Dimensionality Reduction in Hyperspectral Images Using Auto-encoders.** In: International Conference on Advances in Computational Intelligence and Informatics. Springer, Singapore, 2019. p. 101-107.
KENDALL, M. G. **A course in the geometry of n-dimensions**, N. York, Ed. 1961.

LAROSE, Daniel T. **Data mining methods and models.** Hoboken: Wiley-Interscience, 2006.

LAZER, David et al. **The parable of Google Flu: traps in big data analysis**. Science, v. 343, n. 6176, p. 1203-1205, 2014.

LEVADA, Alexandre Luís Magalhães. **Extração de atributos em imagens de sensoriamento remoto utilizando Independent Component Analysis e combinação de métodos lineares.** 2006.

MANLY, Bryan F. J.. **Multivariate statistical methods.** New York, Chapman and Hall, 1986.159 p.

MARQUESONE, Rosangela. **Big Data: Técnicas e tecnologias para extração de valor dos dados.** Editora Casa do Código, 2016.

MUNDOGEO. **Sensoriamento Remoto Hiperespectral.** Disponível em: https://mundogeo.com/2004/08/23/sensoriamento-remoto-hiperespectral/. Acesso em 24 jun. 2021.

PEARSON, Karl. LIII. **On lines and planes of closest fit to systems of points in space.** The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science, v. 2, n. 11, p. 559-572, 1901.

PETEAN, Felipe Coelho de Souza. **Uso de imagens hiperespectrais e da tecnologia LiDAR na identificação de espécies florestais em ambiente urbano na cidade de Belo Horizonte, Minas Gerais**. 2015. Tese de Doutorado. Universidade de São Paulo.

ROSSANT, C. 2015. Disponível em: <https://www.oreilly.com/content/an-illustrated-introduction-to-the-t-sne-algorithm/>.

RŮŽIČKOVÁ, J. A. N. A.; ŠUSTOVÁ, KVĚTOSLAVA. **Determination of selected parameters of quality of the dairy products by NIR spectroscopy.** Czech journal of food sciences, v. 24, p. 255-260, 2006.

SANDANIELO, Vera Lúcia Martins. **"Emprego de técnicas estatísticas na construção de índices de desenvolvimento sustentável aplicados a assentamentos rurais."** (2008): xv-159.

SAQUI, Diego. **Metodologia supervisionada para seleção de bandas de imagens hiperespectrais utilizando o NSGA2.** Tese (Doutorado) - Universidade Federal de São Carlos, January 2018.

SAUL, Lawrence K.; ROWEIS, Sam T. **Think globally, fit locally: unsupervised learning of low dimensional manifolds.** Departmental Papers (CIS), p. 12, 2003.

SINELLI, Nicoletta et al. **Application of near (NIR) infrared and mid (MIR) infrared spectroscopy as a rapid tool to classify extra virgin olive oil on the basis of fruity attribute intensity.** Food research international, v. 43, n. 1, p. 369-375, 2010.

SCOTT, David W. **Multivariate density estimation and visualization.** In: Handbook of computational statistics. Springer, Berlin, Heidelberg, 2012. p. 549-569.

VAN DER MAATEN, Laurens; HINTON, Geoffrey. **Visualizing data using t-SNE.** Journal of machine learning research, v. 9, n. 11, 2008.

VAPNIK, Vladimir; GUYON, Isabel; HASTIE, Trevor. **Support vector machines**. Mach. Learn, v. 20, n. 3, p. 273-297, 1995.

ZHANG, Bing et al. **Hyperspectral image processing and analysis system (HIPAS) and its applications.** Photogrammetric engineering and remote sensing, v. 66, n. 5, p. 605-610, 2000.

Muzambinho, 09 de Julho de 2021