UNIVERSIDADE FEDERAL DO CEARÁ

TECHNICAL REPORT

Aluno: Erick Ramos Coutinho

1. Introdução

O diagnóstico diferencial das doenças "eritematoescamosas" na dermatologia é um desafio devido às semelhanças nas características clínicas apresentadas por essas condições. Doenças como psoríase, dermatite seborreica, líquen plano, pitiríase rósea, dermatite crônica e pitiríase rubra pilar compartilham sintomas de eritema e descamação, dificultando a diferenciação precisa.

Este estudo utiliza uma base de dados que combina informações clínicas e histopatológicas. Os pacientes foram avaliados clinicamente com 12 características e amostras de pele foram coletadas para avaliação de 22 características histopatológicas, analisadas microscópicamente.

O objetivo desta pesquisa é explorar a relação entre as características clínicas e histopatológicas das doenças eritematoescamosas, identificando padrões distintos que possam auxiliar no diagnóstico diferencial. Essa análise detalhada pode fornecer informações valiosas para uma tomada de decisão clínica mais precisa, contribuindo para a seleção de abordagens terapêuticas adequadas e melhorando a qualidade de vida dos pacientes.

2. Observações

Durante a análise dos dados, enfrentamos alguns desafios que precisamos superar para obter resultados confiáveis. Aqui estão alguns desses desafios explicados de forma mais simples:

- ***Compreender as classes:*** Dificuldade para entender o que cada categoria representava nas variáveis do conjunto de dados. Pois elas só são definidas pelos números: 1, 2, 3, 4, 5 e 6
- ***Tratar dados faltantes***: Encontramos alguns valores faltantes na coluna "Age". Para evitar problemas, tivemos que remover esses valores. Assim, garantimos que a análise fosse feita com dados completos e representativos.

  Na análise dos dados, foram utilizadas algumas características para descrever as informações presentes na base de dados. Aqui estão as explicações dessas características de forma simplificada:

- ***Atributo histórico familiar:*** Esse atributo possui o valor 1 se alguma das doenças em questão foi observada na família do paciente e 0 caso contrário. Ele indica se existe um histórico familiar dessas doenças.
- ***Idade:*** Esse recurso simplesmente representa a idade do paciente. É uma informação importante para entender como as doenças podem variar de acordo com a faixa etária.
- ***Características clínicas e histopatológicas:*** Todas as outras características clínicas e histopatológicas receberam um grau entre 0 e 3. Um valor de 0 indica que a característica não estava presente, enquanto 3 indica a maior quantidade possível da característica. Os valores 1 e 2 representam níveis intermediários em relação à presença da característica.

3. Resultados e discussão

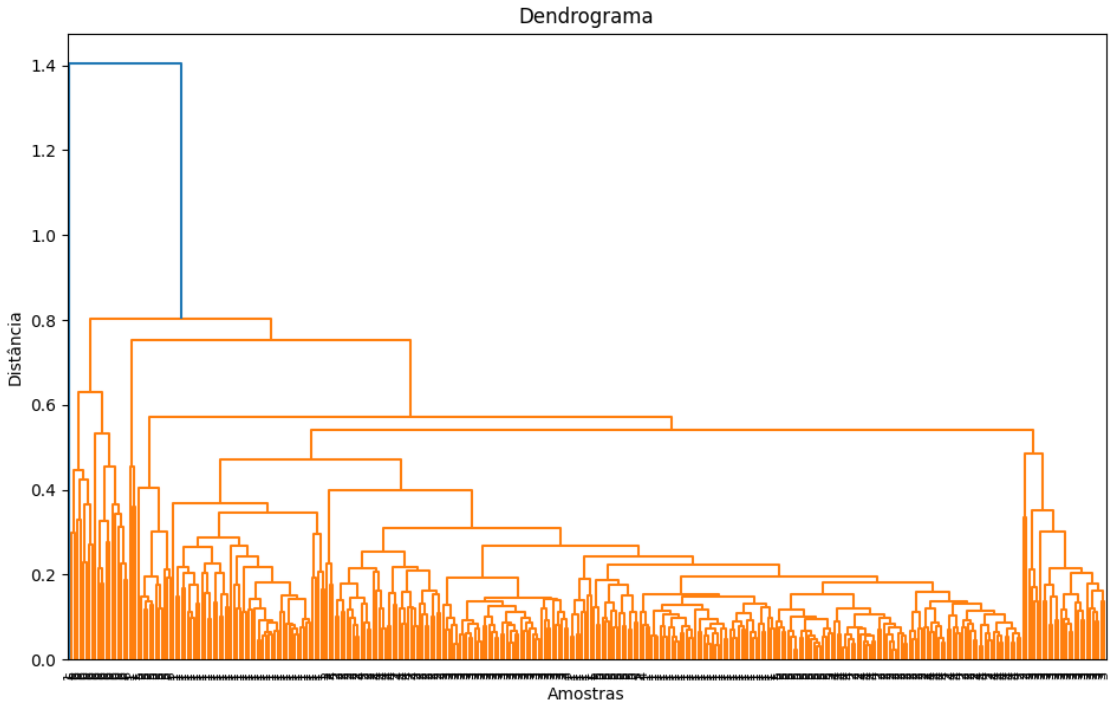
- ***1° Questão -***

Na questão de número 1, foram realizadas as seguintes etapas de pré-processamento e análise dos dados dermatológicos:

Primeiro, as células vazias ou "NaN" na coluna "age" foram identificadas e excluídas, criando um novo data frame chamado "derm1". Em seguida, foi realizado um dendrograma para visualizar a relação entre as amostras e as classes das doenças dermatológicas, utilizando a técnica de linkage completa. Posteriormente, foi aplicado o algoritmo de clusterização K-means para agrupar as amostras em 3 clusters, e um data frame foi criado para mostrar a relação entre os rótulos de cluster e as classes das doenças. Por fim, uma tabela de contingência foi gerada para analisar a distribuição dos rótulos de cluster em relação às classes das doenças.

Resultados Obtidos -

Após a análise do dendrograma abaixo, é perceptível que sua interpretação pode ser desafiadora devido à grande quantidade de dados e à complexidade das informações apresentadas. No entanto, ao examinarmos cuidadosamente o dendrograma em conjunto com as classes do conjunto de dados, foi possível identificar que a estrutura sugere a presença de três grupos distintos. Com base nessa análise, decidimos selecionar o número de clusters igual a 3 para a aplicação do algoritmo de clusterização, visando agrupar as amostras de acordo com suas semelhanças.



Clusterização:

Class Values	1	2	3	4	5	6
Labels						
0	41	25	27	22	19	0
1	31	20	18	17	16	20
2	39	15	26	9	13	0

Com base na tabela de contingência apresentada, podemos observar a distribuição dos rótulos (labels) dos clusters em relação às classes (class_values) das amostras.

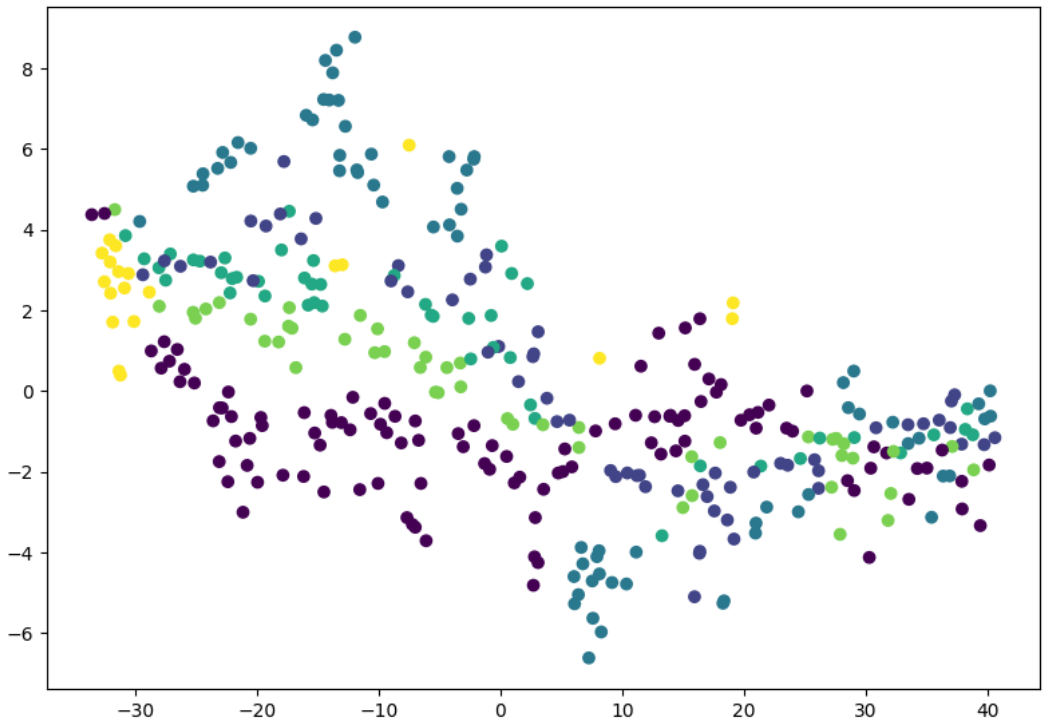
- O cluster 0 possui uma distribuição variada entre as classes, com maior concentração nas classes 1, 2, 3 e 4.
- O cluster 1 também apresenta uma distribuição equilibrada entre as classes, com uma presença significativa nas classes 1, 2, 3 e 4.
- Já o cluster 2 apresenta uma presença dominante na classe 1, com poucas amostras nas demais classes.

Essa análise permite identificar que os clusters formados possuem diferentes padrões de distribuição das classes das doenças dermatológicas, indicando a presença de grupos com características distintas. Essa informação pode ser útil para compreender as semelhanças e diferenças entre as amostras e auxiliar em futuras análises e tomadas de decisão relacionadas às classes das doenças.

● ***2° Questão -***

Na 2° questão, foram realizadas análises de redução de dimensionalidade e visualização dos dados usando as técnicas t-SNE (t-Distributed Stochastic Neighbor Embedding) e PCA (Principal Component Analysis). Essas técnicas permitem representar os dados de forma mais compacta em um espaço bidimensional, facilitando a visualização e compreensão dos padrões e agrupamentos presentes nos dados. No t-SNE, os dados normalizados foram transformados em duas dimensões, enquanto no PCA, os dados padronizados foram projetados em duas dimensões de acordo com as principais componentes. Em ambos os casos, um gráfico de dispersão foi gerado, onde cada ponto representa uma amostra colorida de acordo com a classe. Essas visualizações ajudam a identificar padrões e relações entre as amostras com base em suas características, proporcionando insights valiosos para análises posteriores.

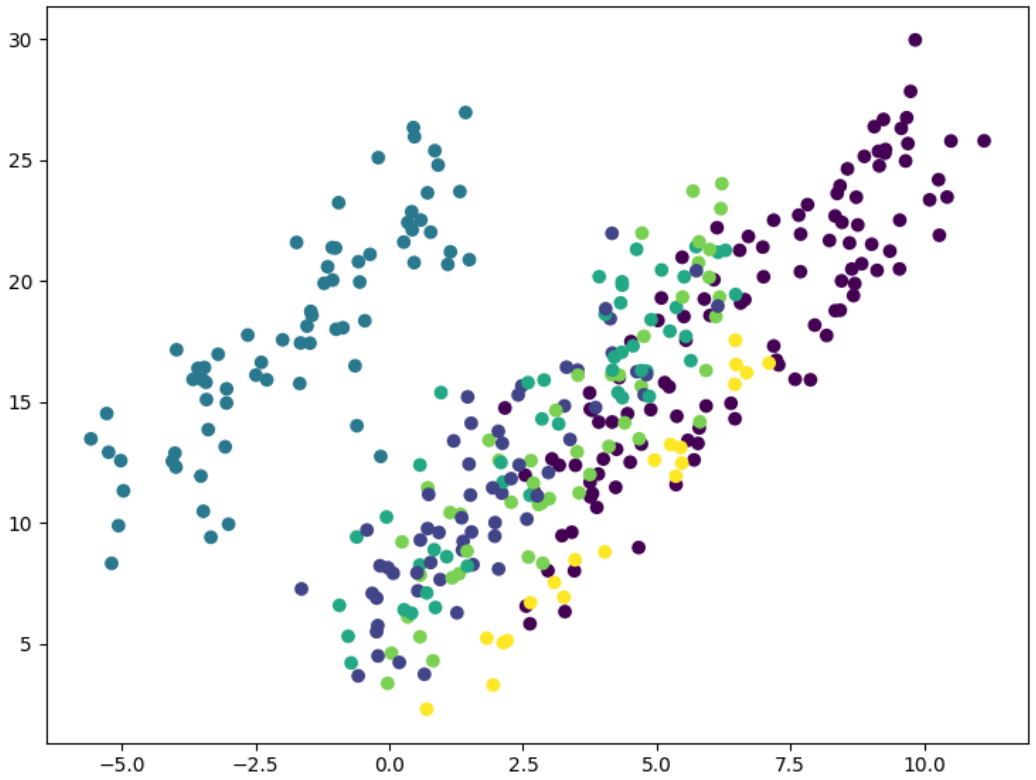
Gráfico de dispersão T-SNE -



O gráfico gerado utilizando a técnica T-SNE (t-Distributed Stochastic Neighbor Embedding) apresenta uma representação visual dos dados em um espaço bidimensional. Cada ponto no gráfico corresponde a uma amostra do conjunto de dados, enquanto as cores dos pontos representam as diferentes classes.

Observando o gráfico, percebemos que os dados estão bastante próximos e as cores não estão claramente separadas, a dispersão e proximidade das cores no gráfico do t-SNE indicam que as amostras das diferentes classes estão misturadas e não formam agrupamentos claros e distintos. Isso pode sugerir que as características presentes nas amostras não são fortemente particulares em relação às classes.

Gráfico de dispersão PCA -



O gráfico do PCA resume as amostras de dados em um espaço bidimensional, com cada ponto representando uma amostra e as cores indicando a classe correspondente. Ele auxilia na identificação de padrões de agrupamento e relações entre as amostras, fornecendo insights sobre a separabilidade das classes.

A análise do gráfico do PCA revela uma maior organização das amostras em relação às cores das classes. Embora ainda haja algumas sobreposições e agrupamentos próximos, as amostras apresentam uma tendência de se agruparem de forma mais distinta em comparação com o gráfico do T-SNE. Além disso, a cor azul está mais separada das demais, indicando uma possível diferenciação mais clara entre essa classe e as outras.

Resultados Obtidos -

Com base na análise dos gráficos do PCA e t-SNE, podemos concluir que as amostras do conjunto de dados possuem uma sobreposição considerável entre as classes. Embora o PCA apresente uma ligeira organização e separação das cores, ainda existe uma conexão significativa entre as amostras. Essa falta de separabilidade entre as classes pode indicar a presença de similaridade nas características das amostras, o que pode dificultar a distinção precisa entre as classes. Portanto, é recomendado explorar outras técnicas de análise, como classificadores mais robustos ou a inclusão de outras variáveis, para obter uma melhor separação entre as classes e uma compreensão mais completa dos padrões presentes nos dados.

- ***3° Questão -***

Na 3° questão, foi realizado o processo de redução de dimensionalidade dos dados utilizando as técnicas de PCA (Análise de Componentes Principais) e t-SNE (t-Distribuição Estocástica de Vizinhança). Em seguida, os dados reduzidos foram divididos em conjuntos de treinamento e teste. Foram criados classificadores k-NN (k-Vizinhos Mais Próximos) para cada conjunto de dados reduzido, e os classificadores foram treinados utilizando os dados de treinamento. Previsões foram feitas nos dados de teste e métricas de avaliação, como relatório de classificação e matriz de confusão, foram calculadas para cada técnica. A acurácia para ambos os métodos foi relatada como 0.33.

- **Métricas de avaliação para PCA -**

  Classification Report:

Class	precision	recall	f1-score	support
1	0.33	0.64	0.44	22
2	0.17	0.14	0.15	14
3	0.43	0.21	0.29	14
4	0.50	0.25	0.33	8
5	0.20	0.08	0.12	12
6	1.00	1.00	1.00	2

accuracy			**0.33**	**72**
macro avg	**0.44**	**0.39**	**0.39**	**72**
weighted avg	**0.33**	**0.33**	**0.30**	**72**

Matriz de confusão para PCA:

14	**3**	**2**	**2**	**1**	**0**
9	**2**	**1**	**0**	**2**	**0**
8	**2**	**3**	**0**	**1**	**0**
5	**1**	**0**	**2**	**0**	**0**
6	**4**	**1**	**0**	**1**	**0**
0	**0**	**0**	**0**	**0**	**2**

- **Métricas de avaliação para t-SNE -**

Classification Report:

Class	precision	recall	f1-score	support
1	**0.33**	**0.59**	**0.42**	**22**
2	**0.18**	**0.14**	**0.16**	**14**
3	**0.55**	**0.43**	**0.48**	**14**
4	**0.33**	**0.12**	**0.18**	**8**
5	**0.50**	**0.25**	**0.33**	**12**
6	**1.00**	**0.50**	**0.67**	**2**

accuracy			**0.33**	**72**
macro avg	**0.48**	**0.34**	**0.37**	**72**
weighted avg	**0.39**	**0.36**	**0.35**	**72**

Matriz de confusão para t-SNE:

13	**3**	**3**	**2**	**1**	**0**
10	**2**	**0**	**0**	**2**	**0**
6	**2**	**6**	**0**	**0**	**0**
5	**1**	**1**	**1**	**0**	**0**
5	**3**	**1**	**0**	**3**	**0**
1	**0**	**0**	**0**	**0**	**1**

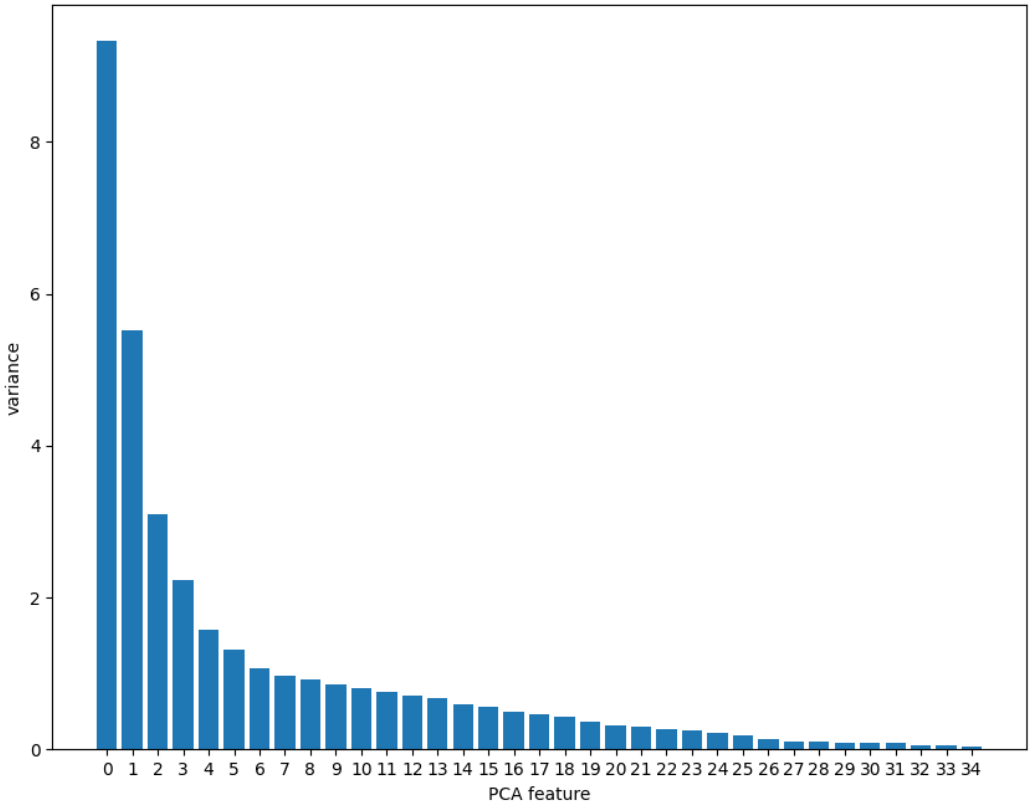
Resultados Obtidos -

Ao analisar as métricas de avaliação para os métodos PCA e t-SNE, podemos observar que ambos apresentam baixa acurácia, em torno de 0.33. No entanto, é importante notar que as métricas de precisão, recall e F1-score para as diferentes classes também são baixas, indicando dificuldades em classificar corretamente as amostras. Além disso, as matrizes de confusão revelam que as classes estão sendo confundidas entre si em ambos os métodos. Portanto, apesar do PCA apresentar uma maior separabilidade das classes no gráfico, os resultados de classificação para ambas as técnicas são limitados, sugerindo a necessidade de explorar outras abordagens ou ajustar os parâmetros dos algoritmos utilizados

- ***4° Questão -***

Na 4° questão é realizada análises de variância usando PCA para identificar as colunas com maior variância no conjunto de dados Dermatology. Em seguida, reduz a dimensionalidade dessas colunas selecionadas usando PCA e t-SNE. Os dados reduzidos (colunas com maior variância) são divididos em conjuntos de treinamento e teste, e classificadores k-NN são treinados com base nesses conjuntos. As previsões são feitas nos dados de teste e são calculadas métricas de avaliação, como precision, recall, F1-score e matriz de confusão, para medir o desempenho dos classificadores. Os resultados, incluindo a acurácia para os métodos PCA (0.64) e t-SNE (0.67), são apresentados.

Gráfico de barras das colunas com maior variância -



As colunas 'erythema', 'scaling', 'definite_borders', 'itching', 'koebner_phenomenon' e 'polygonal_papules' (correspondem aos índices no gráfico acima) apresentaram maior

variância e foram selecionadas para compor um novo dataset. Em seguida, foram aplicadas as métricas de avaliação nesse novo conjunto de dados.

- **Métricas de avaliação para PCA** (colunas com maior variância)**:**

Classification Report:

Class	precision	recall	f1-score	support
1	0.58	0.64	0.61	22
2	0.55	0.43	0.48	14
3	0.87	0.93	0.90	14
4	0.45	0.62	0.53	8
5	0.78	0.58	0.67	12
6	0.50	0.50	0.50	2
accuracy			0.64	72
macro avg	0.62	0.62	0.61	72
weighted avg	0.65	0.64	0.64	72

Matriz de confusão:

14	3	1	4	0	0
5	6	0	1	2	0
1	0	13	0	0	0
1	1	1	5	0	0
2	1	0	1	7	1
1	0	0	0	0	1

- **Métricas de avaliação para t-SNE** (colunas com maior variância) :

Classification Report:

Class	precision	recall	f1-score	support
1	0.62	0.73	0.67	22
2	0.45	0.36	0.40	14
3	1.00	1.00	1.00	14
4	0.44	0.50	0.47	8
5	0.80	0.67	0.73	12
6	0.50	0.50	0.50	2
accuracy			0.67	72
macro avg	0.64	0.63	0.63	72
weighted avg	0.67	0.67	0.66	72

Matriz de confusão:

14	2	0	0	0	2
4	6	0	2	2	1
0	0	14	0	0	0
4	0	0	0	0	0
0	2	1	7	7	0
0	0	0	0	0	2

Resultados Obtidos -

Ao comparar as métricas de avaliação para PCA e t-SNE, observamos que ambos os métodos tiveram um desempenho semelhante em termos de acurácia geral. No entanto, é interessante notar que a seleção das colunas com maior variância pode ter influenciado os resultados.

No caso do PCA, ao utilizar apenas as colunas com maior variância, os resultados mostraram um desempenho relativamente melhor na classe 3, com uma precisão de 87% e recall de 93%. Isso sugere que as informações contidas nessas colunas selecionadas são relevantes para distinguir corretamente a classe 3. Por outro lado, o t-SNE também obteve um desempenho bastante satisfatório para a classe 3, com precisão e recall perfeitos (100%), o que indica que essa classe pode ser facilmente identificada utilizando-se as colunas selecionadas.

O score pode ter subido quando utilizamos apenas as colunas com maior variância em comparação com o uso de todo o dataset devido a algumas razões. Ao selecionar as colunas com maior variância, estamos priorizando as características que possuem uma maior variação nos dados, o que pode indicar que essas características são mais informativas para a classificação. Dessa forma, ao reduzir a dimensionalidade do dataset para apenas essas colunas, estamos concentrando a análise nas características mais relevantes, o que pode levar a um desempenho melhor na tarefa de classificação.

- ***5° Questão -***

Na 5° questão devemos utilizar outra técnica de classificação para comparação com os métodos anteriores, para descobrir o classificador mais adequado. O classificador usado para comparação foi o de Regressão Logística.

- **Métricas de avaliação para Regressão logística:**

Classification Report:

Class	precision	recall	f1-score	support
1	0.77	0.77	0.77	22
2	0.60	0.43	0.50	14

3	**1.00**	**1.00**	**1.00**	**14**
4	**0.64**	**0.88**	**0.74**	**8**
5	**0.73**	**0.67**	**0.70**	**12**
6	**0.25**	**0.50**	**0.33**	**2**
accuracy			**0.74**	**72**
macro avg	**0.66**	**0.71**	**0.67**	**72**
weighted avg	**0.75**	**0.74**	**0.73**	**72**

Matriz de confusão:

17	**1**	**0**	**3**	**0**	**1**
4	**6**	**0**	**0**	**3**	**1**
0	**0**	**14**	**0**	**0**	**0**
1	**0**	**0**	**7**	**0**	**0**
0	**2**	**0**	**1**	**8**	**1**
0	**1**	**0**	**0**	**0**	**1**

Resultados Obtidos -

Ao analisar os resultados da regressão logística, podemos observar que a acurácia obtida foi de 0,74, o que indica que o modelo foi capaz de classificar corretamente 74% das amostras de teste. Ao examinar o relatório de classificação, podemos observar que as classes 1, 3 e 4 tiveram um desempenho relativamente bom, com valores de precisão, recall e F1-score superiores a 0,70. Por outro lado, as classes 2, 5 e 6 apresentaram resultados um pouco mais baixos, com valores inferiores a 0,60 em algumas métricas.

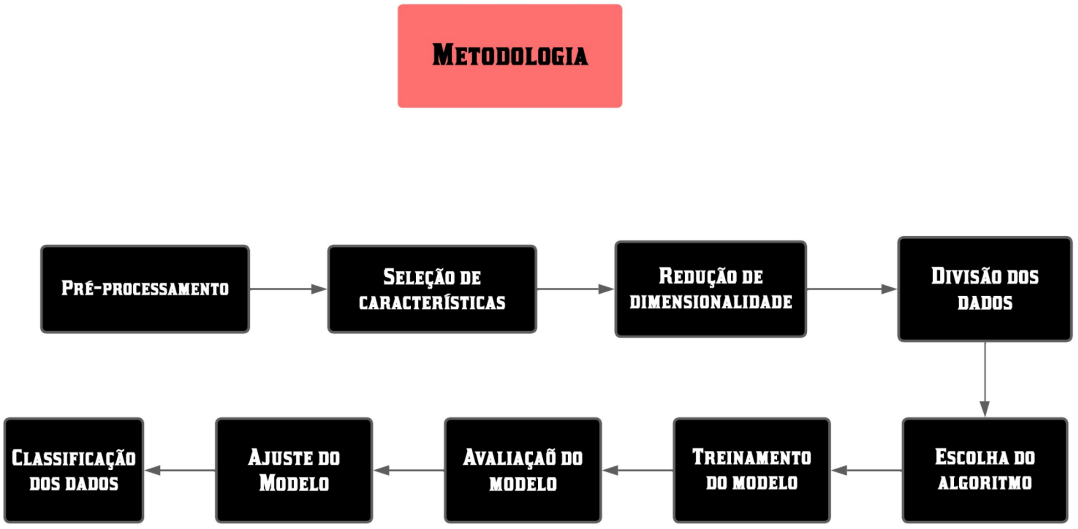
Comparando com os resultados dos métodos anteriores, a regressão logística apresentou uma melhoria significativa em relação ao PCA e ao T-SNE. Isso pode ser atribuído à capacidade da regressão logística de modelar relações mais complexas

entre as variáveis e realizar classificações mais precisas. Em resumo, a regressão logística mostrou um desempenho promissor na classificação do conjunto de dados, alcançando uma acurácia satisfatória.

4. Conclusões

Em conclusão, a análise exploratória e a aplicação de diferentes técnicas de redução de dimensionalidade e classificação revelaram informações valiosas sobre o conjunto de dados dermatológicos. O PCA e o t-SNE permitiram visualizar a distribuição dos dados em espaços de menor dimensão, proporcionando insights sobre a estrutura dos grupos. Embora tenham apresentado limitações na classificação, eles foram úteis para uma compreensão inicial dos dados. Por outro lado, a regressão logística demonstrou um desempenho mais sólido na tarefa de classificação. Isso sugere que a regressão logística é uma abordagem mais adequada para o conjunto de dados estudado, sendo capaz de capturar relações mais complexas entre as variáveis e produzir resultados mais confiáveis.

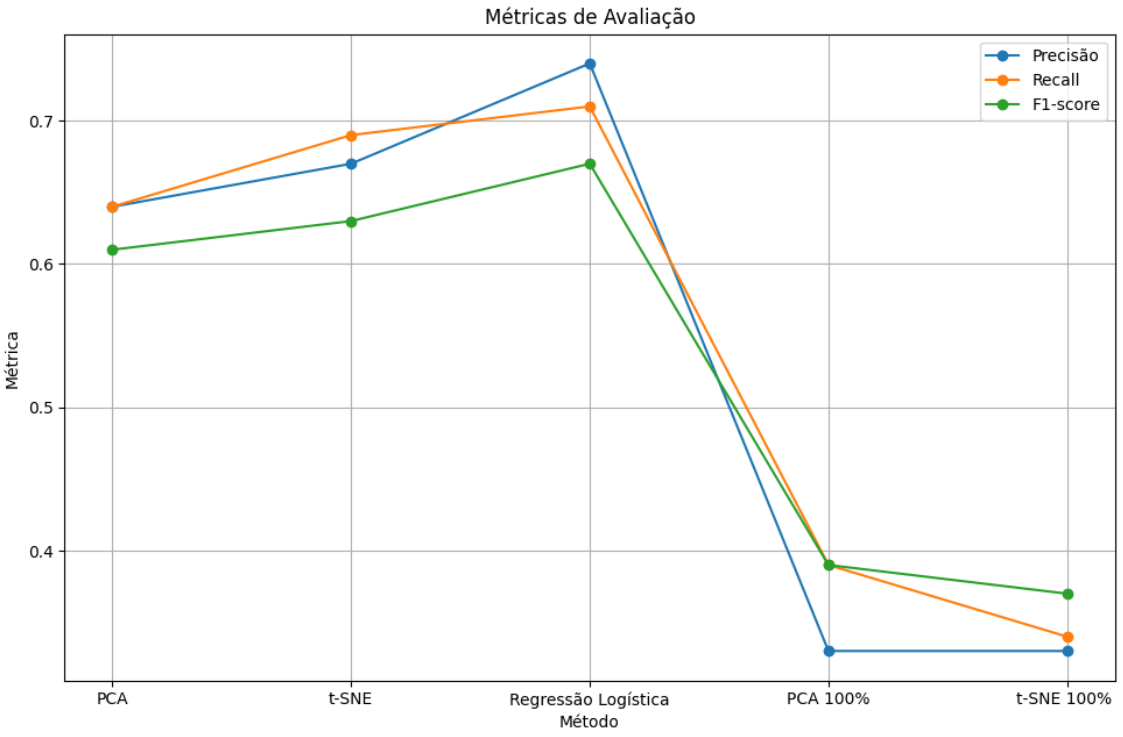
Fluxograma da Metodologia -



Por meio do fluxograma podemos analisar o que foi feito para classificação do dataset de forma organizada e em etapas.

Analisando e comparando os resultados revelou-se que o método de regressão logística obteve uma acurácia de 0.74, superando os resultados anteriores alcançados pelos métodos PCA (0.64) e t-SNE (0.67). Isso indica que a regressão logística foi capaz de fazer previsões mais precisas e consistentes para as classes de doenças dermatológicas. É importante ressaltar que a utilização das colunas com maior variância no conjunto de dados contribuiu para esse resultado mais satisfatório, evidenciando a relevância de selecionar as variáveis mais informativas para a tarefa de classificação. Portanto, essa abordagem combinada de regressão logística e seleção de colunas relevantes apresentou um desempenho superior e promissor no contexto da classificação de doenças dermatológicas.

Gráfico para melhor visualização das métricas -



5. Próximos passos

A. Explorar diferentes algoritmos de classificação:

Além dos algoritmos utilizados neste projeto, como PCA, t-SNE e regressão logística, existem muitos outros algoritmos disponíveis para classificação, como árvores de

decisão, SVM, redes neurais, entre outros. Explore e compare o desempenho desses algoritmos para determinar qual é mais adequado para o seu problema

B. Avaliar e ajustar hiperparâmetros:

Cada algoritmo de classificação possui hiperparâmetros que podem ser ajustados para melhorar seu desempenho. Utilize técnicas de validação cruzada e otimização de hiperparâmetros, como pesquisa em grade ou busca aleatória, para encontrar a combinação ideal de hiperparâmetros que maximize a precisão e o desempenho do modelo.

C. Implementar o modelo em um ambiente de produção:

Uma vez que o modelo tenha sido treinado e validado, é hora de implantá-lo em um ambiente de produção, onde poderá ser usado para fazer previsões em tempo real. Isso pode envolver a integração do modelo em um sistema existente, desenvolvimento de uma interface de usuário amigável ou implantação em um serviço em nuvem.

D. Realizar validação cruzada estratificada:

Ao avaliar o desempenho do modelo, é importante garantir que a validação cruzada seja estratificada, especialmente em conjuntos de dados desequilibrados. Isso ajuda a garantir que todas as classes tenham representação adequada nas divisões de treinamento e teste, evitando resultados tendenciosos.