UNIVERSIDADE FEDERAL DO CEARÁ

TECHNICAL REPORT

Aluno: Larissa Vitória Vasconcelos Sousa

1. Introdução

O conjunto de dados "Financial Distress" disponível no Kaggle e trata-se de um conjunto de dados relacionados à saúde financeira de empresas. O conjunto de dados contém várias variáveis que descrevem as características financeiras e não financeiras de empresas e uma variável alvo chamada "Financial Distress", que indica se a empresa está em dificuldades financeiras ou não. O objetivo é prever se uma empresa está financeiramente em dificuldades ou não. O mesmo possui várias colunas, incluindo informações financeiras e contábeis das empresas. A coluna alvo é "*Financial Distress*", da quarta à última coluna, encontramos as características financeiras e não financeiras da empresa, denotadas por x1 a x83. Essas características pertencem ao período inicial e são usadas para prever se a empresa estará em dificuldades financeiras. Para realizar a análise do conjunto de dados, foi utilizado representação gráfica da hierarquia de agrupamentos (dendrograma), algoritmo de aprendizado não supervisionado *(k-means)*, técnicas de classificação e redução de dimensionalidade utilizando KNN *(K-Nearest Neighbors)*, T-SNE *(t-Distributed Stochastic Neighbor Embedding)* e PCA *(Principal Component Analysis).* Com base nas métricas de desempenho usadas, é possível comparar o desempenho dos dois métodos e determinar qual deles obteve um melhor resultado para o problema de classificação.

2. Observações

Durante o processo de análise, dois problemas surgiram:

1. Normalização: Ao realizar a normalização dos dados na questão 1, foi identificado que existem duas opções diferentes, a normalização por linhas (utilizando a função normalize) e a normalização por colunas (utilizando a função StandardScaler). Após experimentar ambas as abordagens, verificou-se que a normalização por colunas (usando o StandardScaler) resultou em melhores resultados na visualização dos gráficos. Isso indica que essa técnica foi mais adequada para garantir que os dados estivessem na mesma escala e preservar a relação entre os valores em diferentes recursos.
2. Escolha do modelo de classificação: Na questão 5, foi necessário escolher um modelo de classificação. Inicialmente, o modelo de Regressão Logística apresentou uma acurácia de 100%, o que é improvável. Levando isso em consideração, optou-se por utilizar o algoritmo KNN, que obteve uma acurácia

mais realista e plausível. Essa escolha foi feita para evitar resultados excessivamente otimistas que não refletissem a capacidade do modelo de generalizar para dados desconhecidos.

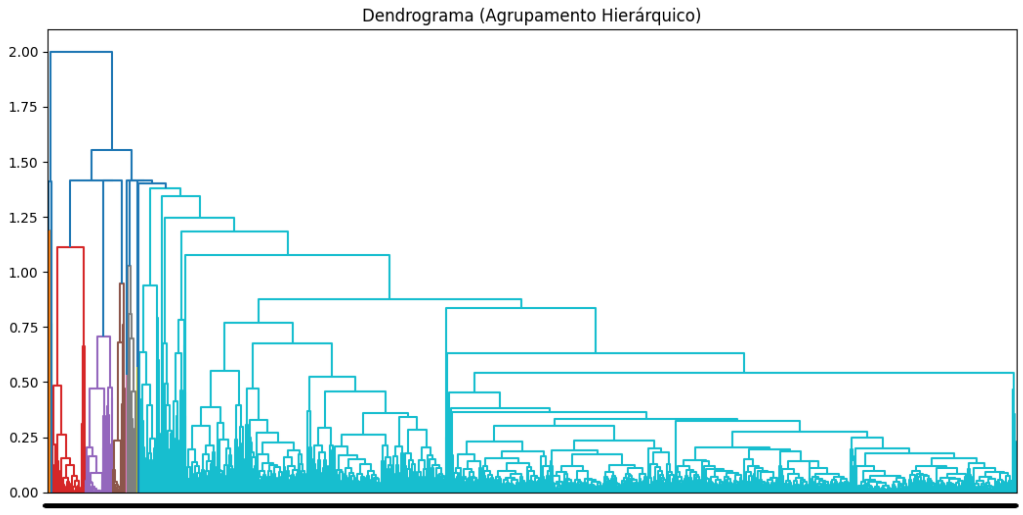
3. Resultados e discussão

Questão 1

Primeiramente, o dataset "*Financial Distress*" é carregado e é feita a transformação da variável alvo "*Financial Distress*" para uma classificação binária: 0 para empresas consideradas saudáveis e 1 para empresas em dificuldades financeiras. Em seguida, é verificado se existem valores vazios ou nulos no *dataset*, e as linhas com esses valores são excluídas. A normalização das linhas dos dados é realizada utilizando a função *normalize*, a fim de garantir que as características estejam na mesma escala e evitar que recursos com escalas diferentes dominem a análise.

O dendrograma (Figura 1) é criado a partir do cálculo do *linkage* utilizando o método "*complete*" e é plotado para visualizar a estrutura hierárquica dos clusters.

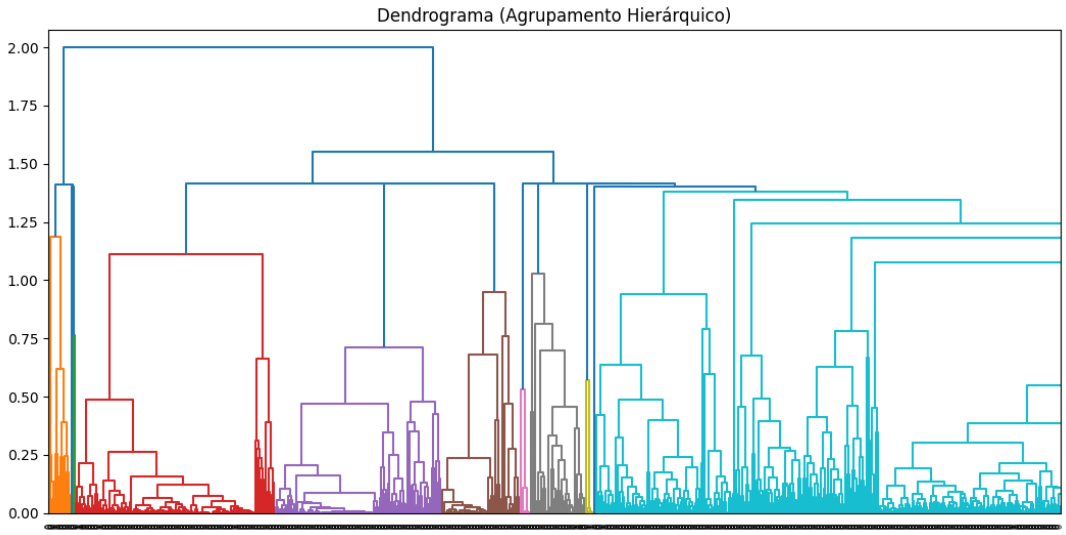
Figura 1 - Dendrograma



Fonte: Autoria própria, 2023.

Na Figura 2, será apresentado o mesmo dendrograma, portanto ampliado, para uma melhor visualização do início da estrutura hierárquica.

Figura 2 – Dendrograma expandido



Fonte: Autoria própria, 2023.

Após a observação do dendrograma, é aplicado o algoritmo k-means com 5 clusters. Em seguida, é gerada uma tabela de contingência (*crosstab*) para analisar a contagem de ocorrências entre os rótulos dos 5 clusters e as categorias da variável alvo “somente_fd” que é somente a coluna “*Financial Distress*”. Essa tabela permite verificar a associação entre as duas variáveis e identificar padrões nos resultados.

Tabela 1 - Crosstab

CROSSTAB			
		SOMENTE_FD (VARIÁVEL ALVO)	
		0	**1**
LABELS	**0**	3526	136
	1	3	0
	2	4	0
	3	2	0
	4	1	0

Fonte: Autoria própria, 2023.

A variável "somente_fd" possui duas categorias: 0 e 1. A variável "labels" também possui diferentes categorias: 0, 1, 2, 3 e 4. A tabela é organizada em formato de matriz, onde as categorias de "somente_fd" são mostradas nas colunas e as categorias de "labels" são mostradas nas linhas.

A interpretação dos valores na tabela é a seguinte:

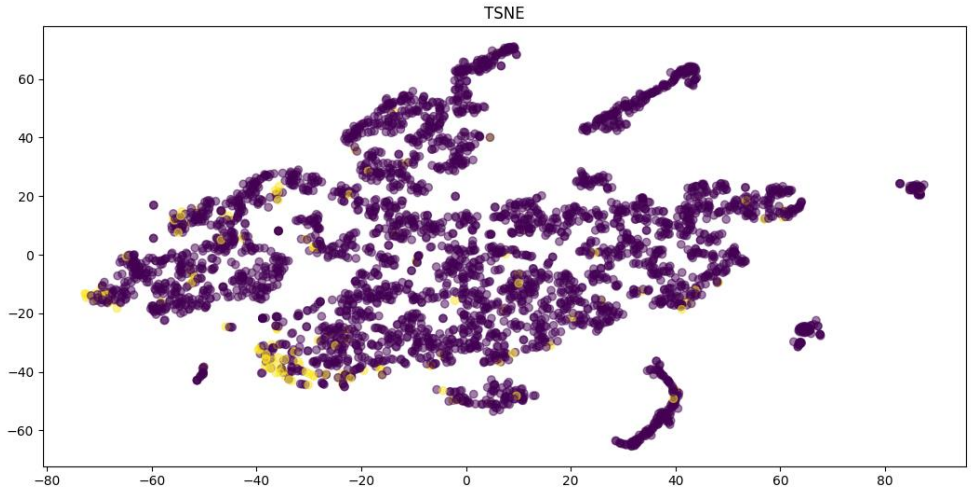
- Na categoria "somente_fd" 0 e "labels" 0, existem 3526 ocorrências;
- Na categoria "somente_fd" 0 e "labels" 1, existem 3 ocorrências;
- Na categoria "somente_fd" 0 e "labels" 2, existem 4 ocorrências;
- Na categoria "somente_fd" 0 e "labels" 3, existem 2 ocorrências;
- Na categoria "somente_fd" 0 e "labels" 4, existe 1 ocorrência;
- Na categoria "somente_fd" 1 e "labels" 0, existem 136 ocorrências;
- Nas demais categorias de "somente_fd" (1, 2, 3, 4) e "labels" (1, 2, 3, 4), não há ocorrências registradas.

Essa análise é útil para identificar grupos de empresas com características semelhantes e entender como esses grupos se relacionam com a variável alvo.

Questão 2

Após realizar a redução de dimensionalidade do *dataset* usando T-SNE e PCA para duas dimensões, foram gerados gráficos de dispersão para visualização das classes, é possível observar que há uma classe (cor: roxo), bem maior que a classe amarela.

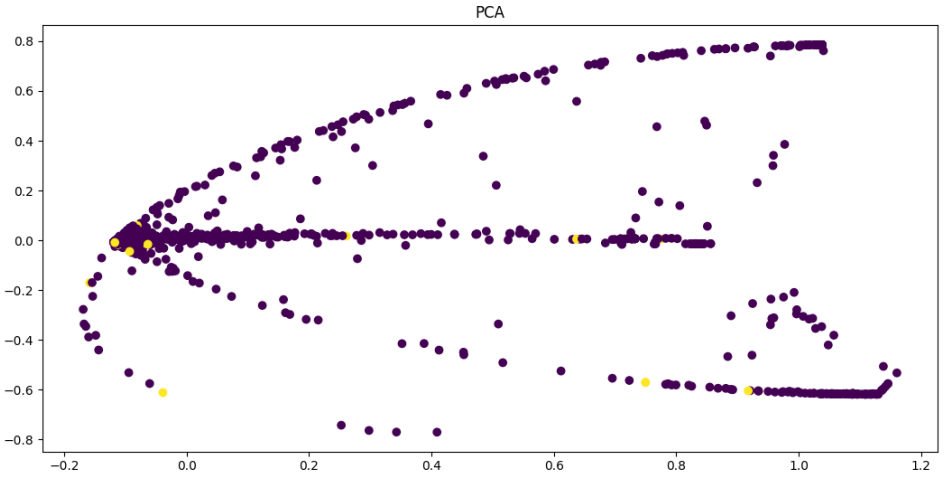
Figura 3 – Redução usando TSNE



Fonte: Autoria própria, 2023.

A Figura 4, irá mostrar a redução de dimensionalidade dos dados usando a técnica PCA (*Principal Component Analysis*).

Figura 4 – Redução usando PCA



Fonte: Autoria própria, 2023.

A seguir, os valores da Tabela 2, foram criados após a aplicação da redução de dimensionalidade do PCA para duas dimensões.

Tabela 2 – Novo dataset com PCA

	LINHAS	COLUNAS
NOVO DATASET PCA	3672	2

Fonte: Autoria própria, 2023.

Os valores indicam que os dados foram transformados em um novo conjunto de dados com 3672 linhas e 2 colunas. A seguir, na Tabela 3, a correlação de Pearson foi calculada.

Tabela 3 - Correlação de Pearson

CORRELAÇÃO
-1.1475734826462863e-16

Fonte: Autoria própria, 2023.

A correlação de Pearson entre as coordenadas x e y é muito próxima de zero. Isso indica que não há uma correlação linear significativa entre as duas dimensões.

Questão 3

Na tabela a seguir, os resultados fornecem as métricas de avaliação e a matriz de confusão para os métodos de PCA e t-SNE, juntamente com as acurácias correspondentes.

Tabela 4 - Métricas de avaliação e a matriz de confusão

Métricas	**Técnicas de redução de dimensionalidade**			
	PCA		**T-SNE**	
	Classe 0	**Classe 1**	**Classe 0**	**Classe 1**
Precision (acurácia geral)	0.94	0	0.95	0.60
Recall	1	0	1	0.07
F1-score	0.97	0	0.97	0.13
Acurácia específica	0.9442176870748299		0.9455782312925171	
Support	694	41	694	41
Matriz de confusão	| 694 0 | | 41 0 |		| 692 2 | | 38 3 |	

Fonte: Autoria própria, 2023.

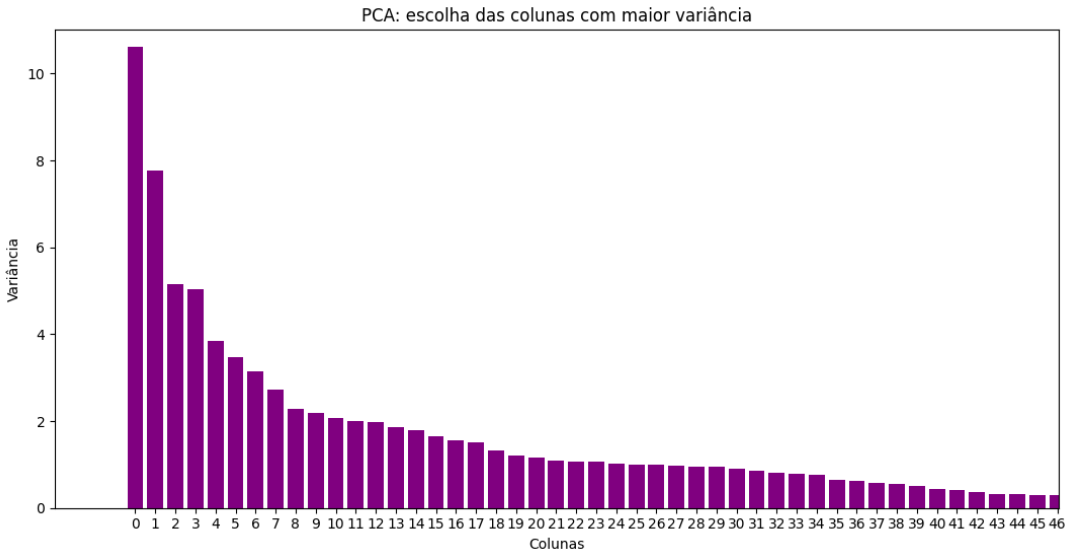
No método PCA, a precisão para a classe 0 é de 0.94, o que significa que 94% das amostras classificadas como classe 0 pelo modelo realmente pertencem à classe 0. Para a classe 1, a precisão é de 0, o que indica que o modelo não conseguiu corretamente prever amostras da classe 1. O recall para a classe 0 é de 1. Isso significa que o modelo identificou corretamente todas as amostras da classe 0. Para a classe 1, o recall é de 0, o que indica que o modelo não conseguiu identificar corretamente nenhuma amostra da classe 1. Sobre a métrica F1-score, a classe 0 é de 0.97, que é uma média harmônica entre a precisão e o recall. Para a classe 1, o F1-score é de 0, indicando um desempenho fraco na previsão dessa classe. O support indica o número de amostras em cada classe, 694 para a classe 0, e 41 para a classe 1. A acurácia geral do modelo é de 0.94, o que significa que ele classificou corretamente 94% das amostras. Sobre a matriz de confusão, mostra que todas as 694 amostras da classe 0 foram corretamente classificadas como classe 0, mas todas as 41 amostras da classe 1 foram incorretamente classificadas como classe 0. Portanto, o modelo não conseguiu identificar nenhuma amostra da classe 1. Por fim, a acurácia específica foi de 0.9442176870748299, ou seja, 94,42%. Do mesmo modo, se interpreta as métricas para a técnica t-SNE, a precisão para a classe 0 é de 0.95, indicando que 95% das amostras classificadas como classe 0 são realmente da

classe 0. Para a classe 1, a precisão é de 0.60, indicando que 60% das amostras classificadas como classe 1 são realmente da classe 1. O recall para a classe 0 é de 1, indicando que o modelo identificou corretamente todas as amostras da classe 0. Para a classe 1, o recall é de 0.07, indicando que apenas 7% das amostras da classe 1 foram corretamente identificadas. O F1-score para a classe 0 é de 0.97, para a classe 1, 0.13, indicando um desempenho muito fraco na previsão dessa classe. O support foi dividido igual para PCA. A acurácia geral do modelo é de 0.95, ou seja, ele classificou corretamente 95% das amostras. A matriz de confusão indica que 692 amostras da classe 0 foram corretamente classificadas como classe 0, enquanto 2 amostras da classe 0 foram incorretamente classificadas como classe 1. Além disso, 38 amostras da classe 1 foram corretamente classificadas, enquanto 3 amostras da classe 1 foram incorretamente classificadas como classe 0. A acurácia específica para o método de t-SNE é de 0.9455782312925171, ou seja, 94,56%. O modelo de PCA não conseguiu identificar corretamente nenhuma amostra da classe 1, enquanto o modelo de t-SNE obteve um desempenho ligeiramente melhor, mas ainda com um recall baixo para a classe 1.

Questão 4

Para essa questão, foi solicitado que utilizasse as colunas com maior variância para os modelos.

Figura 5 – Colunas com maior variância (gráfico ampliado)



Fonte: Autoria própria, 2023.

Foram utilizadas as primeiras 5 colunas (colunas 0 a 4) para realizar as duas técnicas da questão anterior. Vejamos os resultados na Tabela 5:

Tabela 5 - Métricas para as colunas com maior variância

Métricas	**Técnicas de redução de dimensionalidade**			
	PCA		**T-SNE**	
	Classe 0	**Classe 1**	**Classe 0**	**Classe 1**
Precision (acurácia geral)	0.94	0	0.94	0
Recall	1	0	1	0
F1-score	0.97	0	0.97	0
Acurácia específica	0.9401360544217687		0.9414965986394558	
Support	694	41	694	41
Matriz de confusão	| 691 3 | | 41 0 |		| 692 2 | | 41 0 |	

Fonte: Autoria própria, 2023.

É possível notar que tanto para PCA quanto para t-SNE, as métricas de precision, recall e F1-score são iguais para as classes. Isso sugere que ambos os métodos não estão sendo eficazes na classificação da classe 1.

Questão 5

Com base na questão 4, porém utilizando outra técnica de classificação com os mesmos dados, o *K-Nearest Neighbors* (KNN).

Tabela 6 - Métricas para os três classificadores

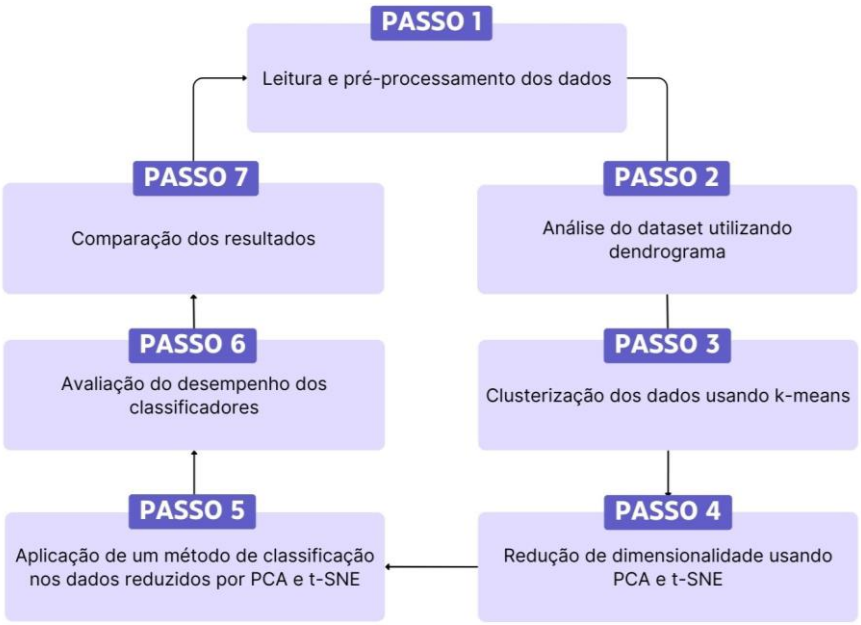
Métricas	Classificadores					
	PCA		T-SNE		KNN	
	Classe 0	Classe 1	Classe 0	Classe 1	Classe 0	Classe 1
Precision (acurácia geral)	0.94	0	0.94	0	0.94	0
Recall	1	0	1	0	1	0
F1-score	0.97	0	0.97	0	0.97	0
Acurácia específica	0.9401360544217687		0.9414965986394558		0.9401360544217687	
Support	694	41	694	41	694	41
Matriz de confusão	| 691 3 | | 41 0 |		| 692 2 | | 41 0 |		| 691 3 | | 41 0 |	

Fonte: Autoria própria, 2023.

Comparando os resultados dos classificadores PCA, t-SNE e KNN, podemos observar que todas as métricas de avaliação são semelhantes entre os três métodos. A precisão (precision), recall, F1-score e acurácia específica são iguais para as classes 0 em todos os classificadores, enquanto todas essas métricas são zero para a classe 1, somente a acurácia específica do t-SNE se sobressai. Isso mostra que nenhum dos classificadores está sendo capaz de distinguir efetivamente a classe 1 dos dados. Portanto, com base nessas métricas, não é possível determinar qual classificador é o melhor.

A metodologia a ser seguida, conforme as questões propostas estará explicada detalhadamente na Figura 5:

Figura 6 – Fluxograma da metodologia



Fonte: Autoria própria, 2023.

Na leitura e pré-processamento dos dados, os dados são lidos, em seguida, é realizado o pré-processamento dos dados, que envolve etapas como remoção de dados ausentes, tratamento de valores inconsistentes, normalização dos dados e codificação de variáveis categóricas, se necessário. O objetivo é preparar os dados para as etapas subsequentes de análise e classificação. No passo 2, ele permite visualizar a similaridade entre os registros e identificar possíveis grupos ou clusters dentro do *dataset*. Essa análise pode fornecer insights sobre a estrutura dos dados e auxiliar na seleção de métodos de clusterização. No terceiro passo, é realizado o processo de clusterização dos dados usando o algoritmo k-means. No passo 4, é realizada a redução de dimensionalidade dos dados utilizando técnicas como Análise de Componentes Principais (PCA) e *t-Distributed Stochastic Neighbor Embedding* (t-SNE). No passo 5, é escolhido um método de classificação, o *K-Nearest Neighbors* (KNN), e é aplicado nos dados reduzidos por PCA e t-SNE. O objetivo é treinar o classificador usando os dados reduzidos e obter um modelo capaz de realizar a classificação de novos registros. No penúltimo passo, é realizada a avaliação do desempenho dos classificadores utilizando métricas apropriadas. Essa avaliação permite verificar o quão bem o classificador está se saindo na tarefa de classificação e pode fornecer informações valiosas para ajustes e melhorias no modelo. E, por fim, no último passo, é feita a comparação dos resultados obtidos.

4. Conclusões

Os resultados obtidos com os classificadores desenvolvidos não atingiram a precisão desejada. Várias razões podem ter contribuído para essa falta de precisão, incluindo a qualidade dos dados utilizados, o tamanho do conjunto de dados, a escolha do classificador, entre outras. A análise das métricas de avaliação dos classificadores PCA, t-SNE e KNN revelou que nenhum deles conseguiu distinguir efetivamente a classe 1 dos dados, apresentando um desempenho fraco. As métricas de precisão, recall, F1-score e acurácia específica foram semelhantes entre os três métodos, com valores adequados para a classe 0, mas todas essas métricas foram zero para a classe 1. Além disso, a comparação entre os métodos de redução de dimensionalidade (PCA e t-SNE) não mostrou grandes diferenças no desempenho dos classificadores. Ambos os métodos forneceram resultados semelhantes em termos de métricas de avaliação. No entanto, o t-SNE obteve uma pequena vantagem na acurácia específica, mas ainda apresentou um recall baixo para a classe 1. Portanto, com base nas métricas avaliadas e nos resultados obtidos, não foi possível determinar com clareza qual classificador é o melhor para este conjunto de dados.

5. Próximos passos

Esses resultados indicam que é necessário um aprimoramento no processo de classificação para melhor identificar e prever corretamente as amostras da classe 1. Uma investigação mais aprofundada deve ser realizada para identificar as possíveis causas desse desempenho insatisfatório, para melhorar a precisão do classificador e buscar resultados melhores. Assim, é necessário a escolha de um outro algoritmo de classificação para melhorar o desempenho e buscar uma solução mais eficaz.