UNIVERSIDADE FEDERAL DO CEARÁ

TECHNICAL REPORT

Aluno: Luciana Sousa Martins

1. Introdução

O conjunto de dados `clinvar_conflicting` usado contém informações sobre variantes genéticas humanas. Existem 65188 observações e 46 variáveis no conjunto de dados, são eles:

- *CHROM*
- *POS*
- *REF*
- *ALT*
- *AF_ESP*
- *AF_EXAC*
- *AF_TGP*
- *CLNDISDB*
- *CLNDISDBINCL*
- *CLNDN*
- *CLNDNINCL*
- *CLNHGVS*
- *CLNSIGINCL*
- *CLNVC*
- *CLNVI*
- *MC*
- *ORIGIN*
- *SSR*
- *CLASS*
- *Allele*
- *Consequence*
- *IMPACT*
- *SYMBOL*
- *Feature_type*
- *Feature*
- *BIOTYPE*
- *EXON*
- *INTRON*
- *cDNA_position*
- *CDS_position*
- *Protein_position*
- *Amino_acids*
- *Codons*
- *Codons*
- *DISTANCE*
- *STRAND*
- *BAM_EDIT*
- *SIFT*
- *PolyPhen*
- *MOTIF_NAME*
- *MOTIF_POS*
- *MOTIF_SCORE_CHANGE*
- *LoFtool*
- *CADD_PHRED*
- *CADD_RAW*
- *BLOSUM6*

2. Observações

Na primeira questão, de início houve conversões em 5 variáveis (‘AF_ESP’, ‘AF_EXAC’, ‘AF_TGP’, ‘CLNVC’, IMPACT’) do tipo ‘string’ e ‘float’ para o tipo ‘int’ e a exclusão de 38 variáveis ('CLNDISDBINCL', 'CLNDNINCL', 'CLNSIGINCL', 'CLNVI', 'SSR', 'INTRON', 'EXON', 'SYMBOL', 'Feature_type', 'Feature', 'BIOTYPE', 'cDNA_position', 'CDS_position',

'Protein_position', 'Amino_acids', 'Codons', 'DISTANCE', 'STRAND', 'BAM_EDIT', 'SIFT', 'PolyPhen', MOTIF_NAME', 'HIGH_INF_POS', 'MOTIF_SCORE_CHANGE', 'LoFtool', 'CADD_PHRED', 'CADD_RAW', 'BLOSUM62', 'MC', 'MOTIF_POS',, 'CLNDISDB', 'CLNDN', 'CLNHGVS', 'REF', 'ALT', 'Allele', 'Consequence', 'CHROM') já que as mesmas tinham valores Nan ou valores do tipo 'string' que não eram tão importantes para a análise.

3. Resultados e discussão

1° Questão:

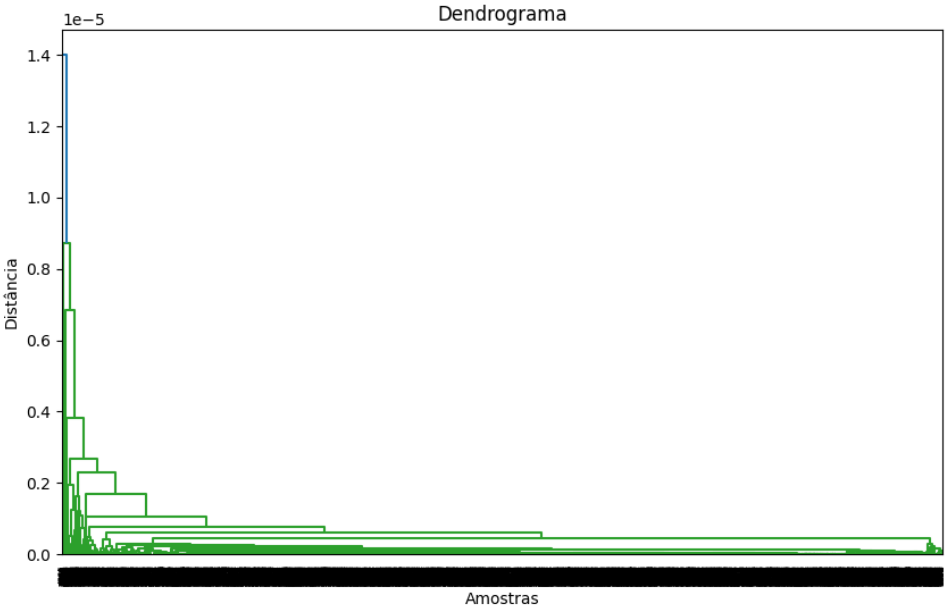
A questão pediu para fazer uma análise do dataset utilizando dendograma (dendograma é um tipo de representação gráfica de dados que é comumente usado em análise de agrupamento (clustering) para mostrar a relação de similaridade entre diferentes objetos), verificando as possibilidades de clusterização e a aplicação do K-means.

O código realizou uma análise das variantes genéticas, foi preciso importar bibliotecas necessárias para geração dos códigos, e seguida criada a variável 'data' e atribuí-la o caminho do dataset. Após uma verificação para identificar células vazias ou NaN e a contagem de valores nulos para cada coluna descobriu-se que a maioria dos atributos da tabela estavam vazios ou do tipo 'string' ou do tipo 'float', sendo assim a exclusão, conversão e substituição de valores por outros valores específicos.

Para a criação de um novo data frame 'df', foi selecionado 10% dos dados do dataframe atualizado; em seguida houve a separação dos dados de entrada (X) das classes (y) do conjunto de dados, a normalização dos dados de entrada usando a função 'normalize' do 'sklearn.preprocessing', o cálculo da ligação entre as amostras usando o método complete e a plotagem do dendograma hierárquico usando a função 'dendogram' do 'scipy.cluster.hierarcy', onde os rótulos das amostras são definidos como as classes y. Em seguida plota novamente o dendograma mas sem os rótulos das amostras, realiza a clusterização usando o algoritmo K-means com 3 clusters e cria um dataframe com os rótulos de clusters e as classes e em seguida imprime a tabela cruzada.

Em resumo, o código realiza a clusterização dos dados usando o algoritmo K-means e visualiza o resultado através do dendograma e da tabela cruzada tendo o objetivo de identificar padrões e agrupamentos nos dados, observando a distribuição das classes em cada cluster.

De acordo com o dendograma abaixo, podemos interpretar que foi criado usando-se uma participação final de apenas 1 agrupamento, de acordo com sua altura nota-se uma maior similaridade entre os elementos agrupados.



Na imagem acima mostra o dendograma da 1° questão

Resultados da Tabela de Frequência Cruzada:

Ela conta a frequência de ocorrência de cada combinação de valores entre as variáveis especificadas e cria uma tabela para exibir essas contagens .

CLASS

0	*1*

Labels

0	*2604*	*889*
1	*793*	*301*
2	*1483*	*449*

Tabela Resultante

labels=0 e CLASS=0	*2604 amostras*
labels=0 e CLASS=1	*889 amostras*

labels=1 e CLASS=0	*793 amostras*
labels=1 e CLASS=1	*301 amostras*
labels=2 e CLASS=0	*1483 amostras*
labels=2 e CLASS=1	*449 amostras*

A tabela resultante mostra a contagem de amostras para cada combinação de rótulos de cluster e classes, permitindo uma análise das correspondências entre eles.

Diante dos resultados acima, há uma distribuição desigual de amostras entre os clusters e classes, já que alguns clusters tem uma quantidade maior de amostras de classes enquanto outros clusters tem uma quantidade mais equilibrada. Os clusters podem ter diferentes níveis de eficácia na separação das classes.

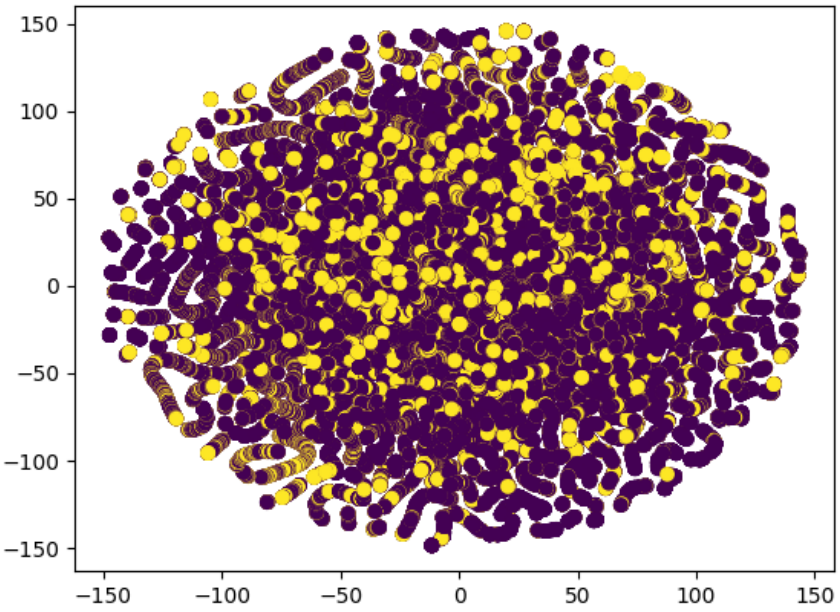
2 Questão:

A questão pediu para reduzir o dataset T-SNE e com PCA para dimensões, plotar o gráfico do atributo que as duas técnicas geraram e de forma subjetiva e visual escolher o gráfico que possui um melhor desempenho em um processo de classificação utilizando os dois atributos.

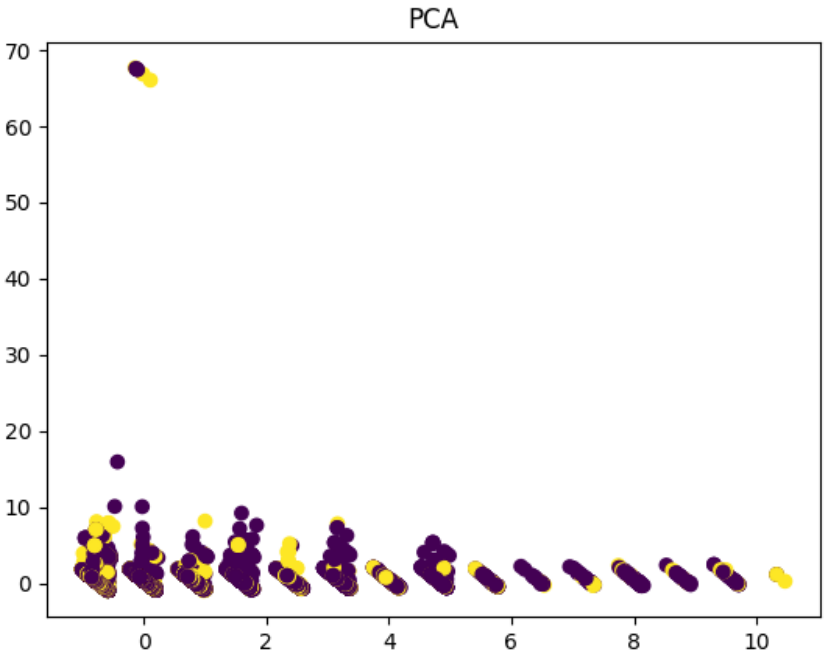
Foi preciso importar bibliotecas necessárias para geração dos códigos, definir as variáveis de entrada e classe a partir do dataframe 'database' e inicializar os dados de entrada usando a função 'normalize' do 'sklearn.proprocessing'. Em seguida foi criado uma instância do algoritmo T-SNE com 2 componentes, uma aplicação do método 'fit_transform' para a obtenção de características T-SNE e selecionada o 0° recurso (dimensão) e 1° recurso (dimensão) das características T-SNE.

Em seguida houve a plotagem de um gráfico de dispersão dos pontos, colorindo de acordo com as classes, cria uma instância do PCA com 2 componentes e ajusta o modelo PCA às amostras escalonadas transformando-as em dimensão reduzida pelo PCA e em seguida exibe o gráfico do PCA.

Em resumo, o código realiza a redução de dimensionalidade dos dados de entrada usando as técnicas T-SNE e PCA e plota gráficos de dispersão para visualização de características reduzidas. Essa análise visual dos gráficos ajuda a avaliar qual das técnicas pode ser mais eficiente em um processo de classificação usando as duas dimensões geradas.



No gráfico acima mostra uma correlação nula já que os pontos não seguem uma tendência positiva nem negativa, significando que não há correlação aparente entre as variáveis.



No gráfico acima mostra uma pequena correlação negativa já que há uma concentração dos pontos em tendência decrescente, ou seja, conforme a variável independente aumenta, a variável dependente diminui.

CLASS

0	*1*

Labels

0	*793*	*301*
1	*2604*	*889*
2	*1483*	*449*

Tabela Resultante

labels=0 e CLASS=0	*793 amostras*
labels=0 e CLASS=1	*301 amostras*
labels=1 e CLASS=0	*2604 amostras*
labels=1 e CLASS=1	*889 amostras*
labels=2 e CLASS=0	*1483 amostras*
labels=2 e CLASS=1	*449 amostras*

A Tabela Resultante mostra a contagem de amostras em cada combinação de labels e classe.

-Para cada combinação de labels e classes, tem uma quantidade de amostras-

3 Questão:

A questão pediu para utilizar os dados da 2° questão para aplicar um método de classificação e para gerar números que quantificam o desempenho do mesmo e em seguida pede para comparar os números classificando o dataset reduzido pelo PCA e T-SNE.

Os dados foram carregados do conjunto de dados original e a dimensionalidade dos dados foi reduzida usando o PCA, com 2 componentes, e o t-SNE, também com 2 componentes, em seguida os dados são divididos em conjuntos de treinamento e teste; em seguida são criados os classificadores K-NN para os dados reduzidos pelo PCA e pelo T-SNE e os mesmo são treinados usando os dados de treinamento.

As métricas de validação são calculadas para cada abordagem de redução de dimensionalidade, as matrizes de confusão são calculadas para cada abordagem e a acurácia é exibida para cada abordagem. Os resultados mostram as métricas de avaliação (precision, recall, f1-score) e as matrizes de confusão para cada abordagem (PCA e t-SNE)

CLASS

0	*1*

Labels

0	*2601*	*888*
1	*794*	*301*
2	*1485*	*450*

Tabela Resultante

labels=0 e CLASS=0	*2601 amostras*
labels=0 e CLASS=1	*888 amostras*
labels=1 e CLASS=0	*749 amostras*
labels=1 e CLASS=1	*301 amostras*
labels=2 e CLASS=0	*1485 amostras*
labels=2 e CLASS=1	*450 amostras*

Métricas de avaliação para PCA

	precision	*recall*	*f1-score*	*support*
0	*0.77*	*0.84*	*0.80*	*9768*
1	*0.33*	*0.23*	*0.27*	*3270*
accuracy			*0.69*	*13038*
macro avg	*0.55*	*0.54*	*0.54*	*13038*
weighted avg	*0.66*	*0.69*	*0.67*	*13038*

Precision: É a proporção de verdadeiros positivos (amostras corretamente classificadas como positivos) em relação ao total de amostras classificadas como positivas. Na tabela, a precisão da classe 0 mostra que 77% das amostras estão corretas e da classe 1 de 33% estão corretas.

Recall: É a proporção de verdadeiros positivos em relação ao total de amostras verdadeiramente positivas. Na tabela, o recall da classe 0 mostra que 84% das amostras foram corretamente identificadas e da classe 1 apenas 23%.

F1-score: É uma medida harmônica entre a precisão e o recall. Na tabela, o F1-score da classe 0 mostra que 80% das amostras tiveram desempenho e da classe 1 apenas 27%.

Support: É o número de amostras verdadeiras para cada classe.

Matriz de Confusão para PCA:

[[8184 1584]
[2505 765]]
Acurácia PCA: 1.00

A matriz de confusão é calculada comparada às classificações feitas pelo modelo com as classes reais dos dados. A matriz de confusão é uma tabela com duas dimensões onde cada linha representa a classe real e cada coluna representa a classe prevista pelo modelo.

Na posição (0,0): 8184 amostras foram classificadas corretamente como pertencentes à classe 0.

Na posição (0,1): 1584 amostras foram classificadas erroneamente como pertencentes à classe 1, mas na verdade pertenciam à classe 0.

Na posição (1,0): 2505 amostras foram classificadas erroneamente como pertencentes à classe 0, mas na verdade pertenciam à classe 1.

Na posição (1,1): 765 amostras foram classificadas corretamente como pertencentes à classe 1.

A acurácia do modelo PCA, é a proporção de amostras corretamente classificadas em relação ao total de amostras. Nesse caso, a acurácia é de 1.00, o que indica que todas as amostras foram classificadas corretamente.

Métricas de avaliação para T-SNE

	precision	*recall*	*f1-score*	*support*
0	*0.77*	*0.84*	*0.80*	*9768*
1	*0.32*	*0.23*	*0.27*	*3270*
accuracy			*0.69*	*13038*
macro avg	*0.54*	*0.53*	*0.53*	*13038*
weighted avg	*0.65*	*0.69*	*0.67*	*13038*

Sobre o resultado das Métricas de avaliação para T-SNE acima:

Precision: Na tabela, a precisão da classe 0 mostra que 77% das amostras estão corretas e da classe 1 de 32% estão corretas.

Recall:. Na tabela, o recall da classe 0 mostra que 84% das amostras foram corretamente identificadas e da classe 1 apenas 23%.

F1-score: Na tabela, o F1-score da classe 0 mostra que 80% das amostras tiveram desempenho e da classe 1 apenas 27%.

Support: É o número de amostras verdadeiras para cada classe.

Matriz de Confusão para T-SNE:

[[8206 1562]
[2520 750]]
Acurácia T-SNE: 1.00

Na posição (0,0): 8206 amostras foram classificadas corretamente como pertencentes à classe 0.

Na posição (0,1): 1562 amostras foram classificadas erroneamente como pertencentes à classe 1, mas na verdade pertenciam à classe 0.

Na posição (1,0): 2520 amostras foram classificadas erroneamente como pertencentes à classe 0, mas na verdade pertenciam à classe 1.

Na posição (1,1): 750 amostras foram classificadas corretamente como pertencentes à classe 1.

A acurácia é de 1.00, o que indica que todas as amostras foram classificadas corretamente.

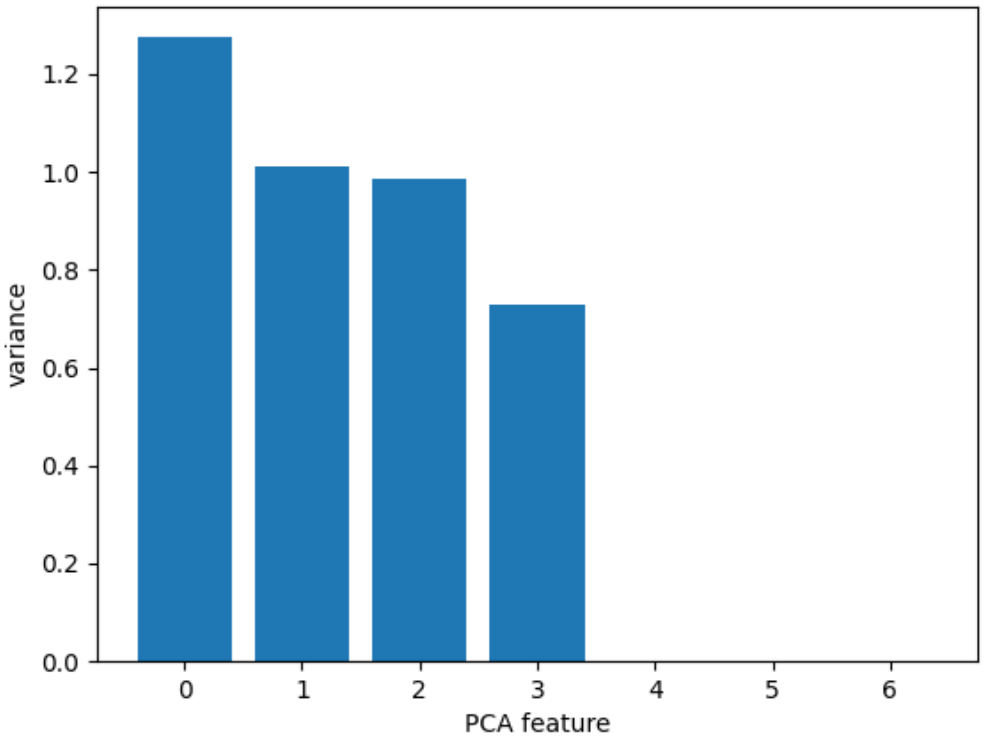
4 Questão:

A questão pediu para utilizar a análise de variância do PCA, reduzir a dimensão para realizar uma classificação utilizando somente as colunas de maior variância. E aplicar o mesmo método de classificação testado na questão 3, e gerar os mesmos números que analisam o desempenho do classificador e verifique se houve melhoria no resultado.

Os dados foram carregados, aplicado o PCA para reduzir a dimensionalidade dos dados e plotado a variância por cada componente principal para selecionar as colunas com maior variância. Em seguida aplicado-se o PCA e T_SNE para redução da dimensionalidade das colunas selecionadas, dividindo os dados reduzidos em conjuntos de treinamento e teste.

Em seguida foi criado e treinado os classificadores K-NN para as abordagens PCA e T-SNE, usando os dados de treinamento. Com as previsões usando os dados de testes, calcula-se as métricas de avaliação e as matrizes de confusão para cada abordagem, e exibe a acurácia para cada abordagem.

Diante do gráfico podemos observar alguns pontos para a análise dos dados, que a variação explicada diminui gradualmente à medida que o número de componentes principais aumenta. Isso ocorre porque os componentes principais subsequentes capturam menos variação em relação aos primeiros.



Na imagem acima é um gráfico mostrando a variação explicada por cada componente principal do PCA.

CLASS

0	*1*

Labels

0	*2604*	*889*
1	*1483*	*449*
2	*793*	*301*

Tabela Resultante

labels=0 e CLASS=0	*2604 amostras*
labels=0 e CLASS=1	*889 amostras*
labels=1 e CLASS=0	*1483 amostras*
labels=1 e CLASS=1	*449 amostras*
labels=2 e CLASS=0	*793 amostras*
labels=2 e CLASS=1	*301 amostras*

Métricas de avaliação para PCA

	precision	*recall*	*f1-score*	*support*
0	*0.77*	*0.84*	*0.80*	*9768*
1	*0.33*	*0.23*	*0.27*	*3270*
accuracy			*0.69*	*13038*
macro avg	*0.55*	*0.54*	*0.54*	*13038*
weighted avg	*0.66*	*0.69*	*0.67*	*13038*

Sobre o resultado das Métricas de avaliação para PCA acima:

Precision: Na tabela, a precisão da classe 0 mostra que 77% das amostras estão corretas e da classe 1 de 33% estão corretas.

Recall:. Na tabela, o recall da classe 0 mostra que 84% das amostras foram corretamente identificadas e da classe 1 apenas 23%.

F1-score: Na tabela, o F1-score da classe 0 mostra que 80% das amostras tiveram desempenho e da classe 1 apenas 27%.

Support: É o número de amostras verdadeiras para cada classe.

Matriz de Confusão para PCA:

[[8191 1577]
[2502 768]]
Acurácia PCA: 1.00

Na posição (0,0): 8191 amostras foram classificadas corretamente como pertencentes à classe 0.

Na posição (0,1): 1577 amostras foram classificadas erroneamente como pertencentes à classe 1, mas na verdade pertenciam à classe 0.

Na posição (1,0): 2502 amostras foram classificadas erroneamente como pertencentes à classe 0, mas na verdade pertenciam à classe 1.

Na posição (1,1): 768 amostras foram classificadas corretamente como pertencentes à classe 1.

A acurácia é de 1.00, o que indica que todas as amostras foram classificadas corretamente.

Métricas de avaliação para T-SNE

	precision	*recall*	*f1-score*	*support*
0	*0.76*	*0.84*	*0.80*	*9768*
1	*0.32*	*0.23*	*0.27*	*3270*
accuracy			*0.69*	*13038*
macro avg	*0.54*	*0.53*	*0.53*	*13038*
weighted avg	*0.65*	*0.69*	*0.67*	*13038*

Sobre o resultado das Métricas de avaliação para T-SNE acima:

Precision: Na tabela, a precisão da classe 0 mostra que 76% das amostras estão corretas e da classe 1 de 32% estão corretas.

Recall:. Na tabela, o recall da classe 0 mostra que 84% das amostras foram corretamente identificadas e da classe 1 apenas 23%.

F1-score: Na tabela, o F1-score da classe 0 mostra que 80% das amostras tiveram desempenho e da classe 1 apenas 27%.

Support: É o número de amostras verdadeiras para cada classe.

Matriz de Confusão para T-SNE:

[[8204 1564]
[2526 744]]
Acurácia T-SNE: 1.00

Na posição (0,0): 8204 amostras foram classificadas corretamente como pertencentes à classe 0.

Na posição (0,1): 1564 amostras foram classificadas erroneamente como pertencentes à classe 1, mas na verdade pertenciam à classe 0.

Na posição (1,0): 2526 amostras foram classificadas erroneamente como pertencentes à classe 0, mas na verdade pertenciam à classe 1.

Na posição (1,1): 744 amostras foram classificadas corretamente como pertencentes à classe 1.

A acurácia é de 1.00, o que indica que todas as amostras foram classificadas corretamente.

5 Questão:

A questão pediu para descobrir qual classificador mais adequado fazendo uma comparação entre classificadores, e utilizar outra técnica de classificadões com os mesmo dados para gerar os números que quantificam o desempenho para fazer uma comparação entre eles.

Primeiramente, o conjunto de dados é preparado, dividindo-se em atributos de entrada (X) e classes de destino (y). Em seguida, as colunas com maior variância são selecionadas, em seguida o PCA e o t-SNE são aplicados aos atributos de entrada com o objetivo de reduzir a dimensionalidade dos dados para 2 componentes, os dados são divididos em conjuntos de treinamento e teste e em seguida são criados classificadores k-NN (k-Nearest Neighbors) para o PCA e o t-SNE, e esses classificadores são treinados

com os dados de treinamento. Com as previsões feitas nos dados de teste usando os classificadores de treino, as métricas de avaliação são calculadas para o PCA e T-SNE, e também é exibida a matriz de confusão e a acurácia .

Em seguida, é utilizado outro classificador, a Regressão Logística, para realizar a classificação nos dados com as colunas de maior variância. O classificador é treinado, as previsões são feitas e é gerada a matriz de confusão. Finalmente, é gerado o relatório de classificação com as métricas de avaliação e é calculada a acurácia da Regressão Logística

CLASS

0	*1*

Labels

0	*1483*	*449*
1	*2604*	*889*
2	*793*	*301*

Tabela Resultante

labels=0 e CLASS=0	*1483 amostras*
labels=0 e CLASS=1	*449 amostras*
labels=1 e CLASS=0	*2604 amostras*
labels=1 e CLASS=1	*889 amostras*
labels=2 e CLASS=0	*793 amostras*
labels=2 e CLASS=1	*301 amostras*

Métricas de avaliação para PCA

	precision	*recall*	*f1-score*	*support*
0	*0.77*	*0.84*	*0.80*	*9768*
1	*0.33*	*0.23*	*0.27*	*3270*
accuracy			*0.69*	*13038*
macro avg	*0.55*	*0.54*	*0.54*	*13038*
weighted avg	*0.66*	*0.69*	*0.67*	*13038*

Sobre o resultado das Métricas de avaliação para PCA acima:

Precision: Na tabela, a precisão da classe 0 mostra que 77% das amostras estão corretas e da classe 1 de 33% estão corretas.

Recall:. Na tabela, o recall da classe 0 mostra que 84% das amostras foram corretamente identificadas e da classe 1 apenas 23%.

F1-score: Na tabela, o F1-score da classe 0 mostra que 80% das amostras tiveram desempenho e da classe 1 apenas 27%.

Support: É o número de amostras verdadeiras para cada classe.

Matriz de Confusão para PCA:

[[8190 1578]
[2502 768]]
Acurácia PCA: 1.00

Na posição (0,0): 8190 amostras foram classificadas corretamente como pertencentes à classe 0.

Na posição (0,1): 1578 amostras foram classificadas erroneamente como pertencentes à classe 1, mas na verdade pertenciam à classe 0.

Na posição (1,0): 2502 amostras foram classificadas erroneamente como pertencentes à classe 0, mas na verdade pertenciam à classe 1.

Na posição (1,1): 768 amostras foram classificadas corretamente como pertencentes à classe 1.

A acurácia é de 1.00, o que indica que todas as amostras foram classificadas corretamente.

Métricas de avaliação para T-SNE

	precision	*recall*	*f1-score*	*support*
0	*0.77*	*0.84*	*0.80*	*9768*
1	*0.32*	*0.23*	*0.27*	*3270*
accuracy			*0.69*	*13038*
macro avg	*0.54*	*0.53*	*0.53*	*13038*
weighted avg	*0.65*	*0.69*	*0.67*	*13038*

Sobre o resultado das Métricas de avaliação para T-SNE acima:

Precision: Na tabela, a precisão da classe 0 mostra que 77% das amostras estão corretas e da classe 1 de 32% estão corretas.

Recall:. Na tabela, o recall da classe 0 mostra que 84% das amostras foram corretamente identificadas e da classe 1 apenas 23%.

F1-score: Na tabela, o F1-score da classe 0 mostra que 80% das amostras tiveram desempenho e da classe 1 apenas 27%.

Support: É o número de amostras verdadeiras para cada classe.

Matriz de Confusão para T-SNE:

[[8205 1563]
[2519 751]
Acurácia T-SNE: 1.00

Na posição (0,0): 8205 amostras foram classificadas corretamente como pertencentes à classe 0.

Na posição (0,1): 1563 amostras foram classificadas erroneamente como pertencentes à classe 1, mas na verdade pertenciam à classe 0.

Na posição (1,0): 2519 amostras foram classificadas erroneamente como pertencentes à classe 0, mas na verdade pertenciam à classe 1.

Na posição (1,1): 751 amostras foram classificadas corretamente como pertencentes à classe 1.

A acurácia é de 1.00, o que indica que todas as amostras foram classificadas corretamente.

Classificador: Regressão Logística

Matriz de confusão:
[[9768 0]
[3270 0]]

A matriz de confusão apresentada indica os resultados do classificador de regressão logística.

A matriz é composta por quatro elementos: verdadeiros positivos (9768), falsos positivos (0), falsos negativos (3270) e verdadeiros negativos (0).

Esses resultados indicam que o classificador de regressão logística não está sendo capaz de capturar a classe positiva em seu modelo. É possível que haja algum desequilíbrio na distribuição das classes ou que o modelo não esteja sendo capaz de encontrar um padrão adequado para separar as classes.

Relatório de Classificação:

	precision	*recall*	*f1-score*	*support*
0	*0.75*	*1.00*	*0.86*	*9768*
1	*0.00*	*0.00*	*0.00*	*3270*
accuracy			*0.75*	*13038*
macro avg	*0.37*	*0.50*	*0.43*	*13038*
weighted avg	*0.56*	*0.75*	*0.64*	*13038*

No caso apresentado, a classe 0 possui uma alta precisão de 75%, o que indica que a maioria das amostras classificadas como 0 são realmente da classe 0. No entanto, a classe 1 possui uma precisão de 0%, o que indica que todas as amostras classificadas como 1 estão incorretas.

O recall (revocação) para a classe 0 é de 100%, o que significa que todas as amostras da classe 0 foram corretamente identificadas. No entanto, o recall para a classe 1 é de 0%, indicando que nenhuma das amostras da classe 1 foi corretamente identificada.

O F1-score é uma métrica que combina a precisão e o recall em uma única medida. Para a classe 0, o F1-score é de 86%, indicando um bom equilíbrio entre precisão e recall. Já para a classe 1, o F1-score é de 0%, indicando um desempenho muito baixo na identificação dessa classe.

A acurácia geral do classificador é de 0.75, o que significa que 75% das amostras foram classificadas corretamente. No entanto, é importante observar que a acurácia pode ser enganosa quando as classes estão desequilibradas, como é o caso aqui.

Acurácia da Regressão Logística: 0.75

A acurácia da Regressão Logística é de 0.75, o que significa que o modelo classificou corretamente 75% das amostras do conjunto de dados.

4. Conclusões

Os resultados não foram satisfatórios, pois diante da análise feita, podemos concluir que o gráfico de dispersão com dimensão reduzida mostrou uma relação nula e negativa entre as variáveis independente e dependente, mostrando que as variáveis reduzidas não estão fortemente relacionadas ou correlacionadas com a variável dependente. Já as tabelas de resultados para as Métricas de avaliação para T-SNE e PCA mostram um bom desempenho das métricas em relação a classe 0 e um baixo desempenho em relação a classe 1, Isso sugere que o modelo ou método de classificação utilizado pode estar tendo dificuldades em distinguir corretamente os exemplos da classe 1.

5. Próximos passos

Para uma melhor análise e precisão de resultados para interpretação:

É importante investigar as causas desse baixo desempenho e considerar possíveis estratégias de melhoria, como ajustar os hiperparâmetros do modelo ou utilizar técnicas de balanceamento de classes. No geral, é necessário analisar mais detalhadamente as características dos dados, a escolha do modelo de classificação e as métricas de avaliação para ter uma compreensão mais completa do desempenho do classificador e das relações entre as variáveis. Além disso, é recomendado considerar a aplicação de outras técnicas de análise exploratória e visualização de dados para obter insights adicionais.

Avaliar outras métricas de avaliação: Além das métricas utilizadas até agora (precisão, recall, f1-score, acurácia), considere outras métricas de avaliação, como curva ROC, área sob a curva ROC, sensibilidade, especificidade, entre outras, dependendo das necessidades específicas do projeto.

Explorar outras técnicas de redução de dimensionalidade: Além do PCA e do t-SNE, você pode explorar outras técnicas de redução de dimensionalidade, como LDA (Análise de Discriminante Linear) ou Autoencoders, para verificar se elas podem fornecer melhores resultados ou insights adicionais sobre os dados.

link para visualização de códigos:

https://github.com/alysonbz/IA/tree/Luciana_Martins/AV3