UNIVERSIDADE ESTADUAL DE CAMPINAS

Bernardo Abib, 236053

Redução de dimensionalidade e K-vizinhos mais próximos na predição de diagnósticos de tumores

Campinas - SP

2024

Sumário

Resumo

O objetivo deste trabalho é construir um modelo de classificação de aprendizado de máquina supervisionado que realize predições do diagnóstico de tumores baseando-se em dados sobre suas medidas físicas.

Para isso, foi construído um modelo de k-vizinhos mais próximos por distância euclidiana das 30 variáveis numéricas do conjunto de dados. Este modelo obteve alta acurácia nas predições, mesmo trabalhando com dados de alta dimensão.

Afim de otimizar o modelo, realizou-se uma análise de componentes principais, cujos componentes selecionados foram utilizados de preditores/features em um novo modelo de k-vizinhos mais próximos, agora operando em consideravelmente menos dimensões do que o anterior.

1. Análises iniciais

1.1 Descrição dos dados

São 568 observações de tumores classificados como benignos ou malígnos, e dez características físicas destes. Cada característica tem média, desvio padrão e maior medida como informações separadas. É lógico assumir, então, que as variáveis são altamente correlacionadas.

As características contidas no banco de dados são raio, textura, perímetro, área, suavidade, compacidade, concavidade, pontos côncavos, simetria e dimensão fractal. Como cada uma delas dá origem a 3 variáveis diferentes (média, desvio padrão e maior medida), o conjunto de dados tem um total de 31 variáveis, contando com o diagnóstico.

1.2 Análise exploratória

Abaixo, compara-se os boxplots de cada uma das variáveis, separando as observações por diagnóstico. Vale observar que as porcentagens de tumor benigno/maligno são 62.7/37.3, respectivamente.

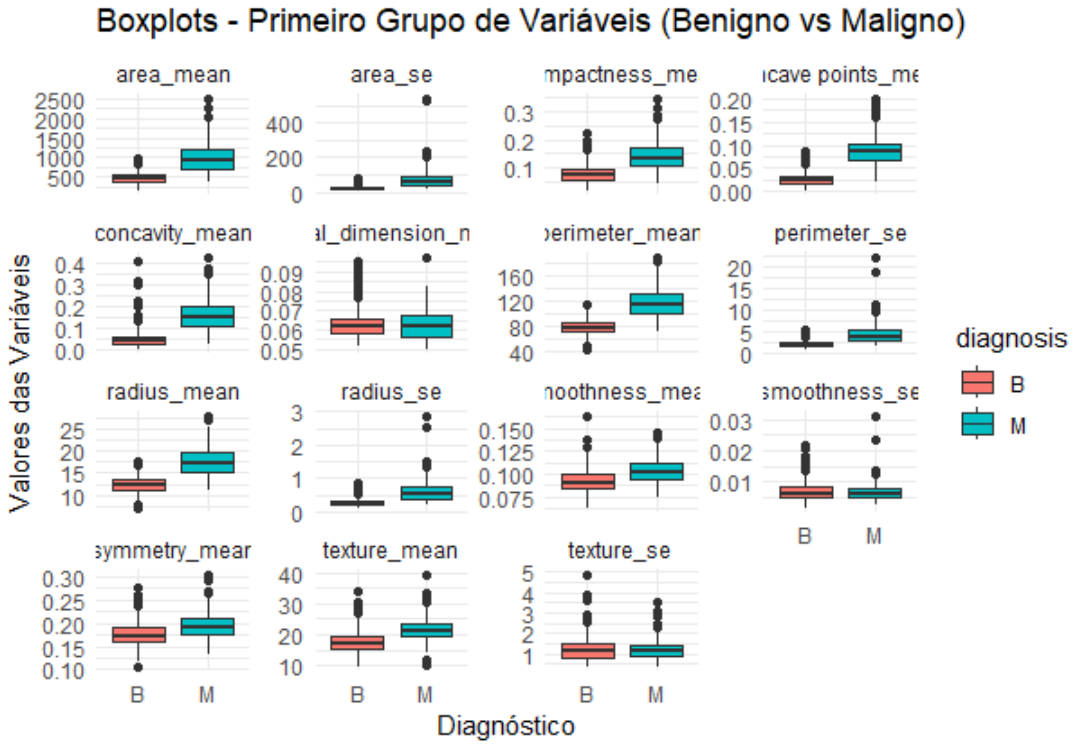

Figura 1 –

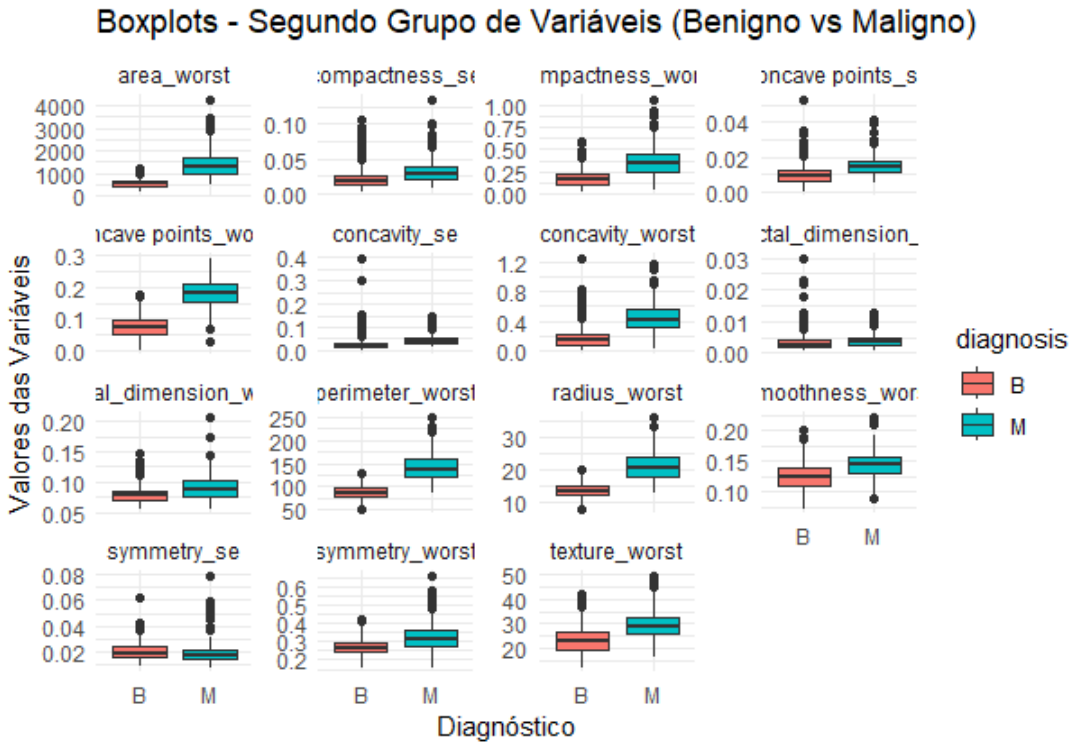

Figura 2 –

Por se tratarem de dados de alta dimensão, as medidas descritivas específicas de cada variável não são interessantes. As comparações entre os grupos, no entanto, podem indicar as variáveis que contém maior discrepância - uma provável diferença estatisticamente significativa - entre tumores benignos e malignos.

Nos boxplots, é possível observar ao menos 8 variáveis onde o primeiro quartil das observações de tumores malígnos é maior do que o terceiro quartil das observações de tumores benígnos, indicando que observações de diagnóstico malígno podem ser significativamente maiores em múltiplas medidas.

Podemos observar melhor esses padrões no gráfico abaixo:

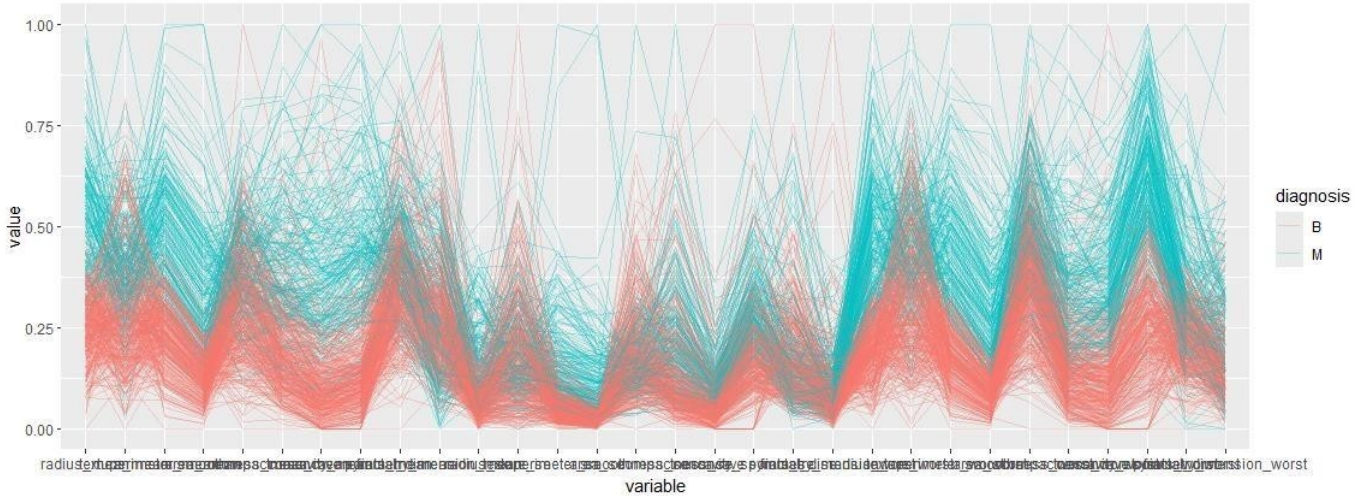

Figura 3 – Coordenadas paralelas

Apesar dos nomes das variáveis não estarem legíveis, o gráfico de coordenadas paralelas torna mais visível o padrão intuído na análise dos boxplots. É possível notar que, na maior parte dos eixos (que indicam cada uma das variáveis), as linhas azuis (observações malígnas) parecem se agrupar acima das linhas vermelhas (observações benignas), com padrões nitidamente distintos para cada tipo de diagnóstico.

É importante também mencionar a alta correlação entre as variáveis. Como discutido anteriormente, cada medida do tumor dá origem a 3 variáveis diferentes. Além disso, muitas das medidas são naturalmente relacionadas (como por exemplo, o raio e a área do tumor).

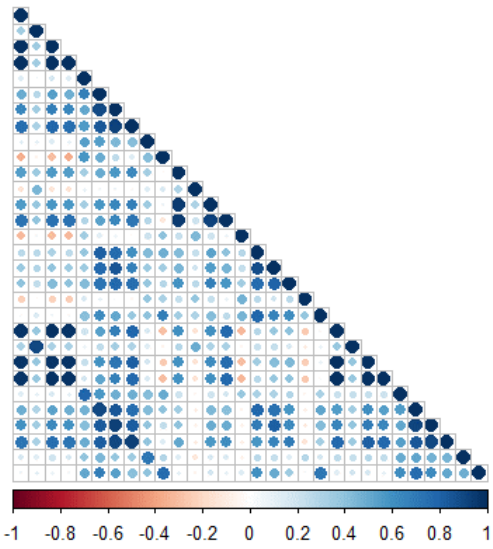

Figura 4 – Matriz de correlação de Spearman

2. Modelo de classificação com os dados originais

2.1 Preparação dos dados

Como modelos KNN baseiam-se na distância dos pontos dos preditores/features para a classificação de uma observação, variáveis em escalas diferentes (com valores ou amplitudes muito maiores do que outras) podem ofuscar variáveis relevantes que estejam numa escala menor, e afetar drasticamente a precisão das predições. Por isso, antes do ajuste do modelo, as variáveis numéricas foram normalizadas - desta forma, eliminando o impacto que diferentes escalas/unidades poderiam ter na distância dos valores dos preditores.

Após a normalização das variáveis numéricas, dividiu-se o conjunto de dados em dois subgrupos: 3/4 das observações foram selecionados para treinar o modelo, e 1/4 foi separado para a realização de testes após o ajuste.

2.2 Ajuste do K-vizinhos mais próximos

Para decidir a quantidade de vizinhos a serem selecionados para a classificação de uma observação, comparou-se a taxa de erro de cada cenário:

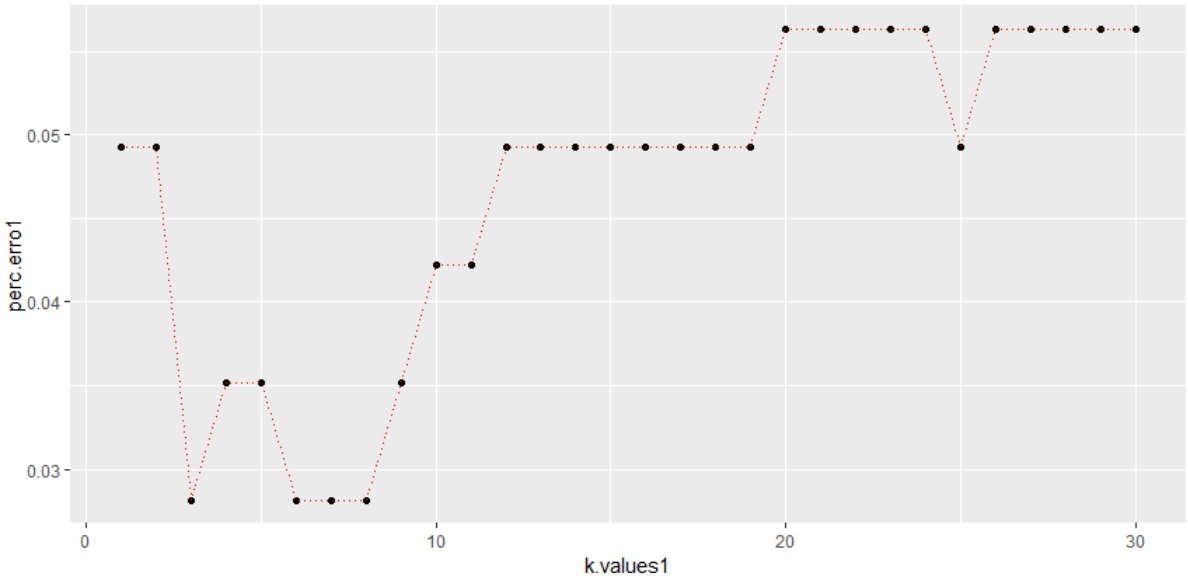

Figura 5 – Erro por quantidade de vizinhos

Observa-se que o menor erro percentual na classificação das observações acontece quando considera-se as 3 observações vizinhas mais próximas. Com essa escolha, o modelo obteve 97,18 porcento de precisão na classificação das observações-teste: todos os 89 tumores benignos foram dianosticados corretamente, enquanto a predição de tumores malignos acertou 49/53. Ou seja, a predição de diagnóstico de 132 tumores contou com apenas 4 erros.

3. Análise de Componentes Principais

Como discutido na análise exploratória, as variáveis numéricas do conjunto de dados possuem altíssima correlação entre si, e são relacionadas por natureza. É coerente supor, portanto, que muitas dessas medidas sejam redundantes, e que talvez não sejam essenciais para a representação da variabilidade dos dados.

Por isso, realizou-se uma redução de dimensionalidade dos dados através da análise de componentes principais. Desta forma, deixa-se de trabalhar com 30 variáveis numéricas, e passa-se a usar combinações lineares das variáveis originais, cujos coeficientes são os elementos dos autovetores associados à matriz de correlação. A quantidade dessas combinações lineares - chamadas de componentes - é igual a quantidade de variáveis originais. Porém, esses componentes são não correlacionados, e estão em ordem decrescente de explicação da variabilidade total dos dados.

Após a aplicação da ACP, observa-se o impacto dos primeiros componentes:

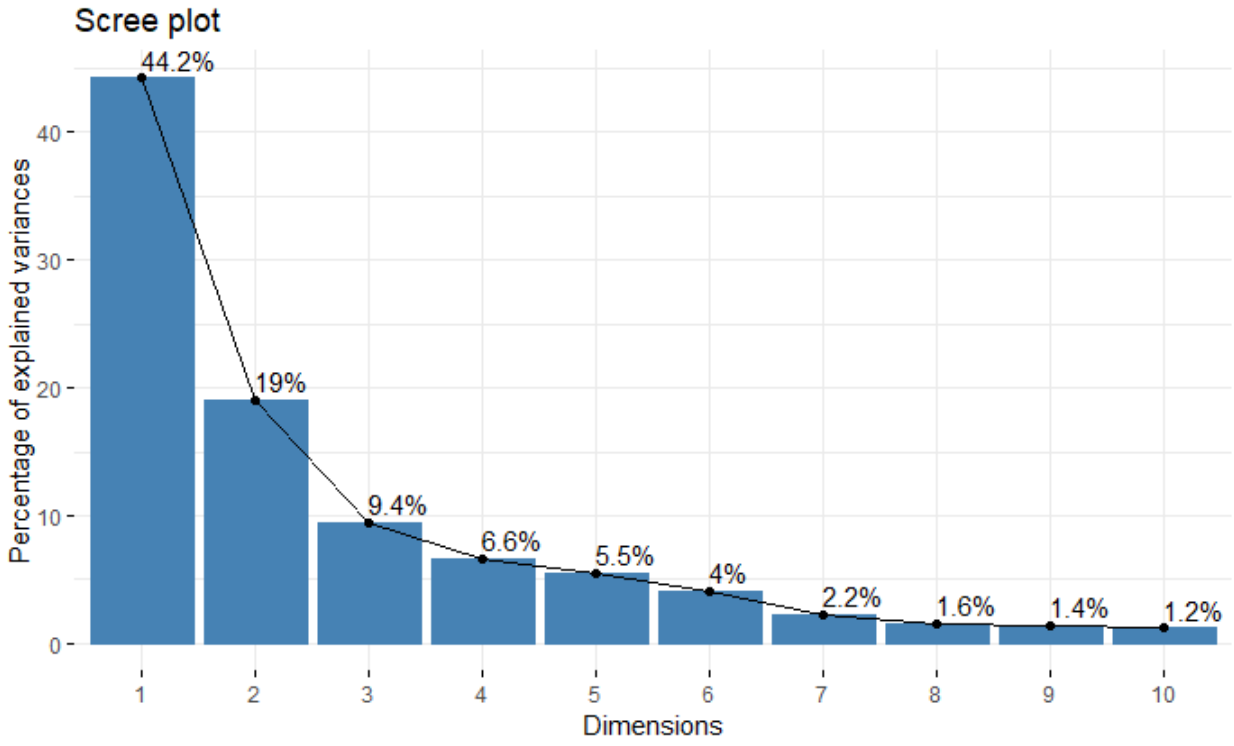

Figura 6 – Explicação da variância por componente

Nota-se que com os 10 componentes principais - 20 dimensões a menos que os dados originais - explica-se 95 porcento da variância total. E com apenas os dois componentes principais, explica-se mais de 60 porcento.

4. Modelo de classificação usando componentes principais

Agora, com os componentes principais carregando grande parte da informação dos dados, é viável reduzir drasticamente a dimensionalidade em que nosso modelo de k-vizinhos mais próximos opera. Para isso, deve-se escolher um número de componentes que faça sentido.

Primeiramente, testou-se o modelo com os seis componentes principais - quantidade escolhida pelo critério de Kaisser - e observou-se o seguinte resutado:

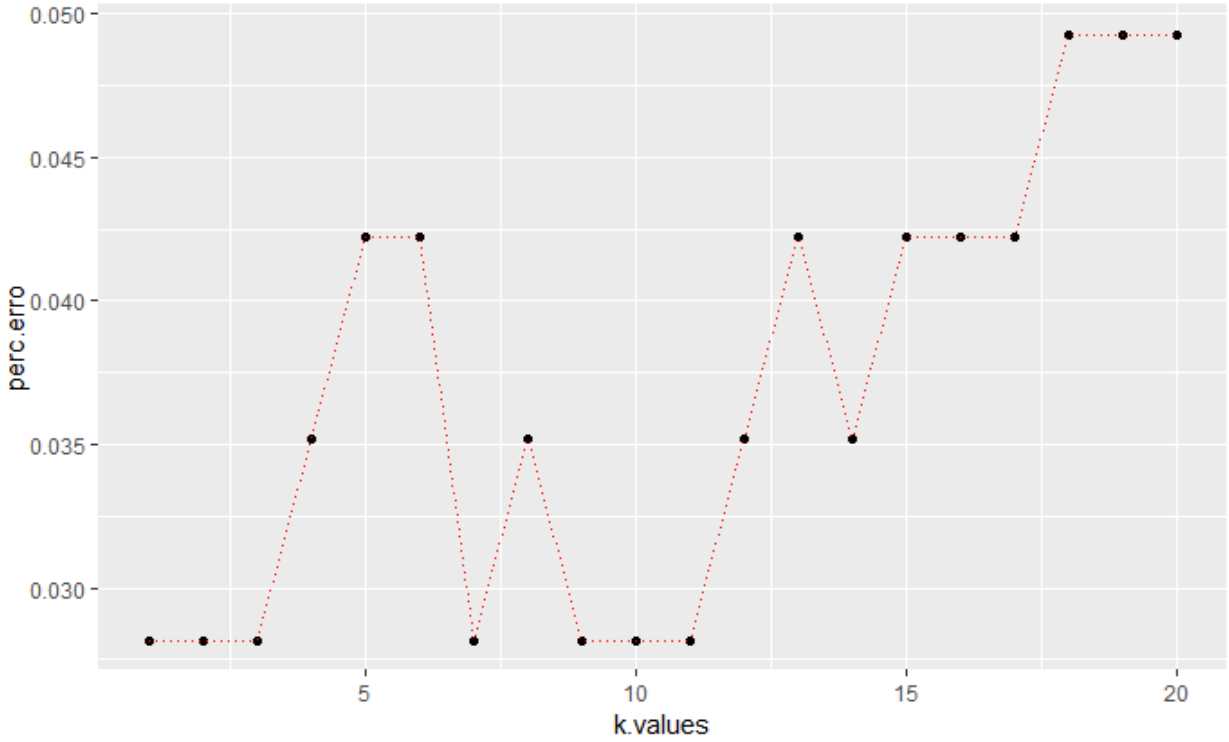

Figura 7 – erro por quantidade de vizinhos (6 componentes)

Desta forma, usando o vizinho mais próximo para a classificação do diagnóstico do tumor, obteve-se uma acurácia de 97,18 porcento - desempenho idêntico ao modelo com os dados iniciais, que operava em 24 dimensões adicionais.

Nesse cenário, faz sentido buscar um número ótimo de componentes que reduza ao máximo a dimensionalidade dos dados, mantendo a precisão do modelo. Abaixo, um gráfico do erro percentual por quantidade de vizinhos considerados, separado por quantidade de componentes principais selecionados:

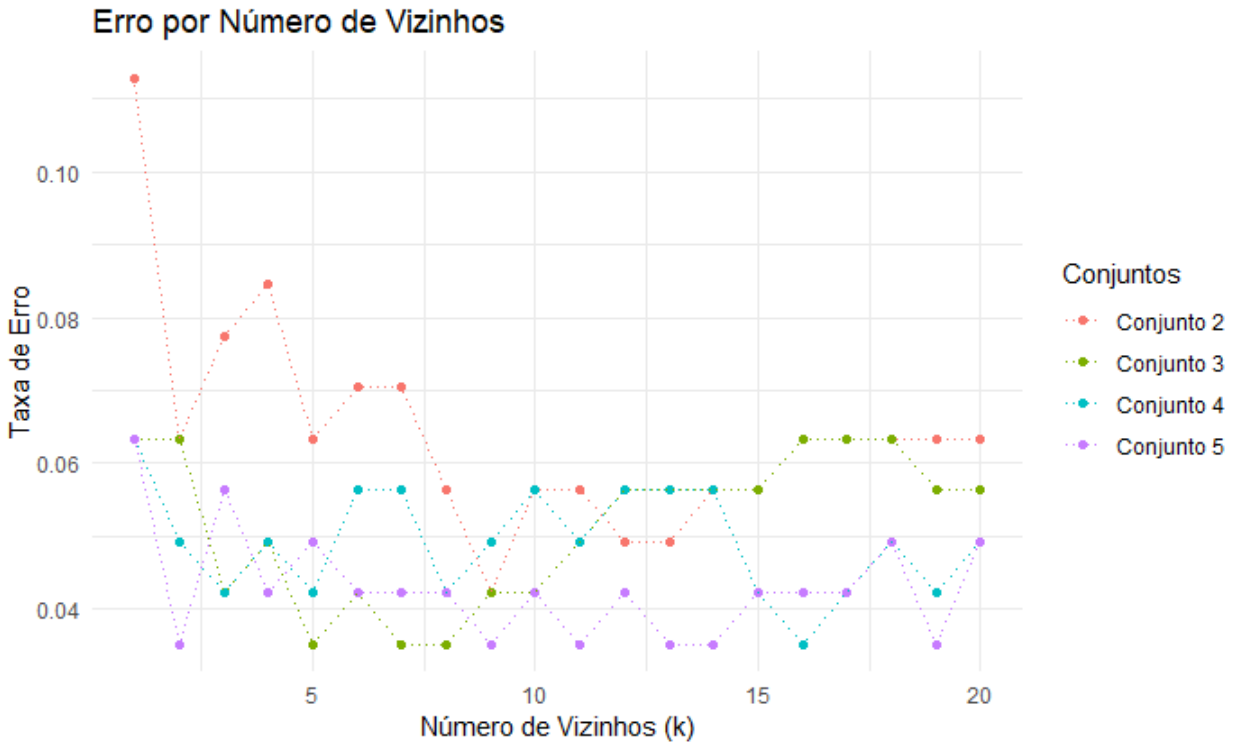

Figura 8 – erro por quantidade de vizinhos (2 a 5 componentes)

Nota-se que, usando o conjunto 2 (contendo apenas os 2 componentes principais), é possível obter menos de 0,05 de erro ao usar os 9 vizinhos mais próximos. Nota-se também que todos os conjuntos 3, 4 e 5 (contendo respectivamente os 3, 4 e 5 componentes principais) atingem a precisão máxima de 96,48 porcento em determinado número de vizinhos considerados, porém nenhum atingiu a precisão de 97,18 porcento obtida com os dados originais e utilizando os 6 componentes principais.

5. Conclusão

Inicialmente, foi construído um modelo de classificação que operava sob os dados originais, em 30 dimensões. Após a análise de componentes principais, foi estudada a performance de modelos de acordo com a quantidade de componentes selecionados, e concluiu-se que o KNN com 6 componentes principais considerando o primeiro vizinho mais próximo, é o modelo de menor dimensão a manter a exata precisão do modelo anterior, com 97,18 porcento de acurácia na predição de diagnóstico de tumores (89/89 para benignos e 49/53 para malignos).

Outro modelo que chama atenção, é o KNN com os 3 componentes principais considerando os 5 vizinhos mais próximos, que obteve 96,48 porcento de acurácia nos diagnósticos (89/89 para benignos e 48/53 para malignos), obtendo uma precisão quase identica ao modelo original, porém com 27 dimensões a menos.

É importante reconhecer que, com a ACP, há uma perda na interpretabilidade do modelo, visto que não estão mais sendo trabalhadas as variáveis originais. No entanto, a significativa redução de dimensionalidade atrelada a preservação da alta acurácia das predições, justifica por completo o uso dos componentes principais como preditores do modelo de classificação.

6. Bibliografia

- Johnson, R. A., Wichern, D. W. (2007). "Applied multivariate statistical analysis". 6ª ed. Capítulo 4
- Härdle, W. K., Simar, L. (2019). "Applied Multivariate Statistical Analysis". 5ª ed. Capítulos 7 e 11
- POLOWCZYK, Agnieszka; POLOWCZYK, Alicja. The Effectiveness of PCA in KNN, Gaussian Naive Bayes Classifier and SVM for Raisin Dataset. 2021
- Luc Devroye, Laszl ´ o Gy ´ orfi, and G ¨ abor Lugosi. ´ A probabilistic theory of pattern recognition, volume 31. Springer Science Business Media, 1996.