Teoria do Aprendizado Estatístico

Atividade 3

Luiz Henrique Barretta Francisco - 202100155302

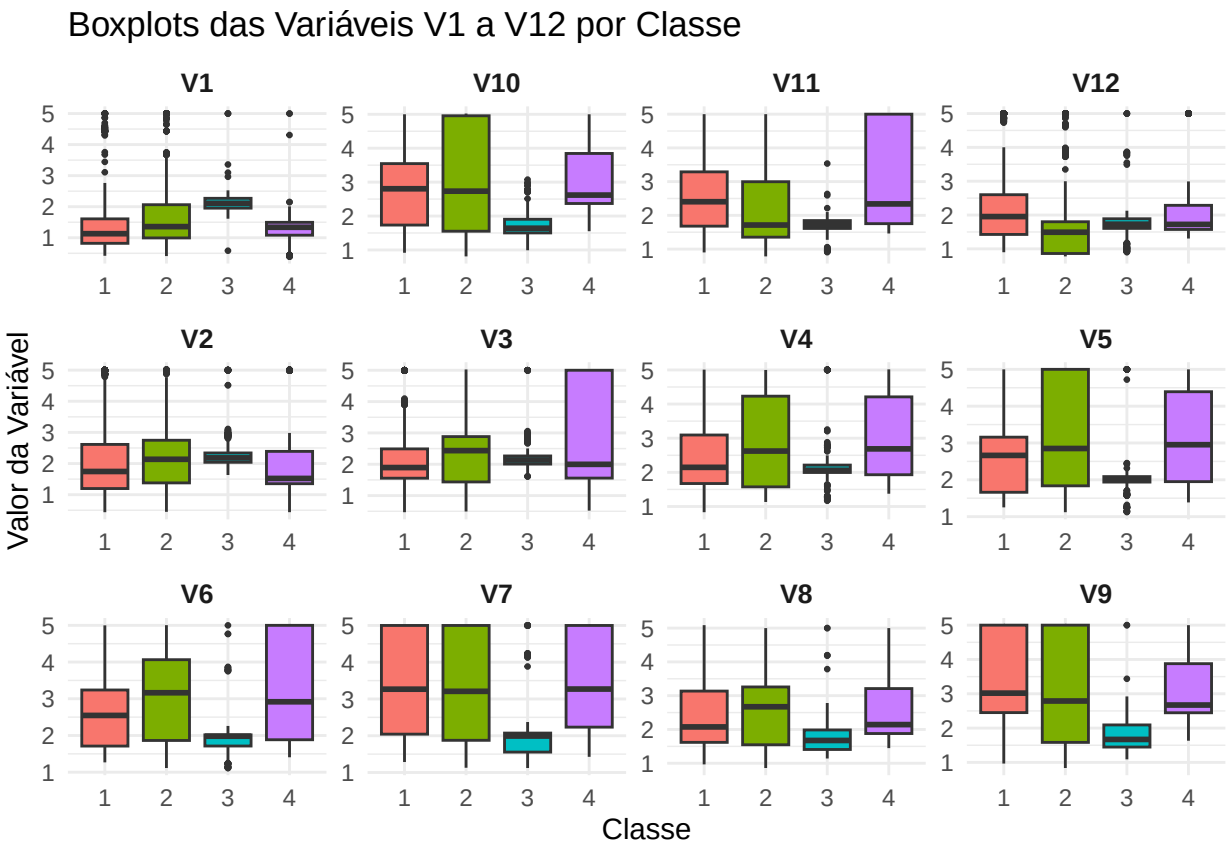
maio/2025

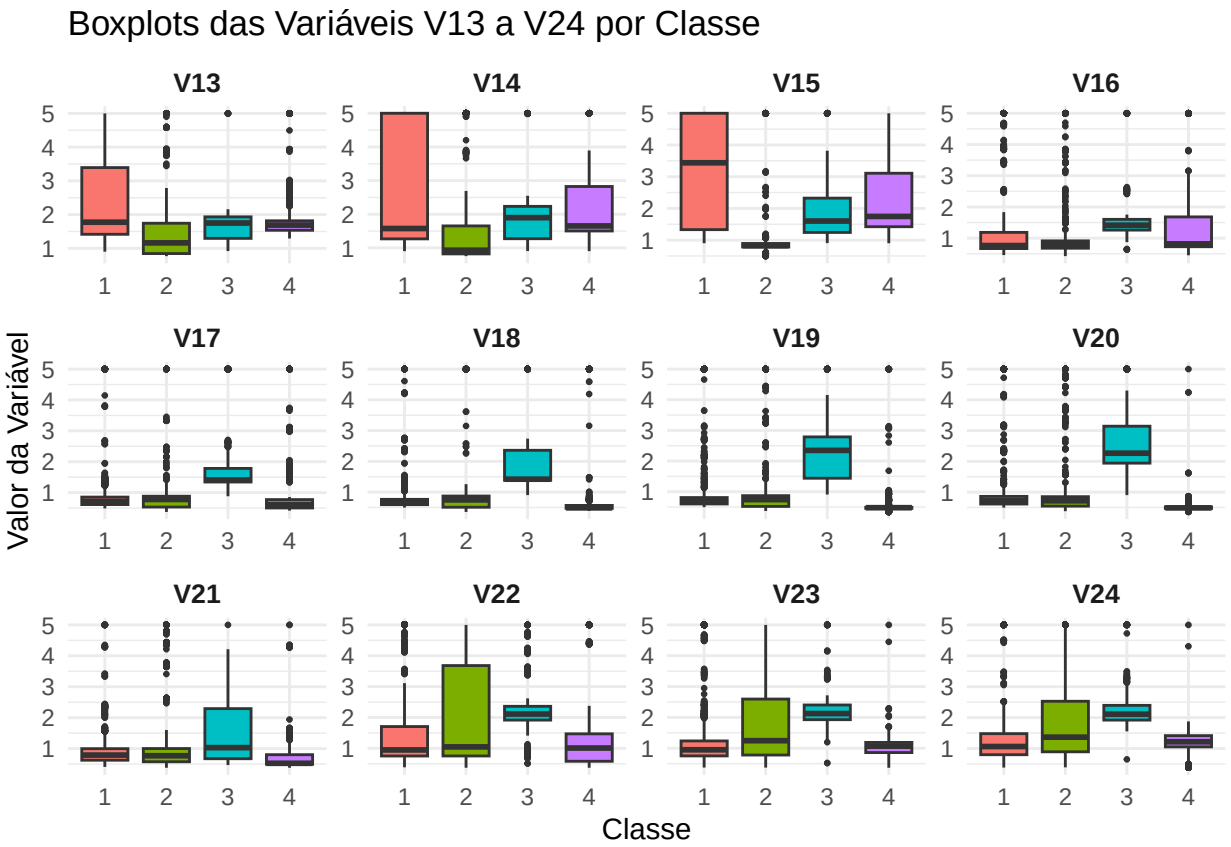
Introdução

Este trabalho tem como objetivo aplicar e comparar dois métodos clássicos de classificação supervisionada — o k-Nearest Neighbors (KNN) e a Regressão Logística — utilizando o conjunto de dados "Robo.csv". O foco será a avaliação do desempenho desses modelos por meio de técnicas de validação cruzada, análise de medidas de desempenho e visualização das fronteiras de decisão. Além disso, serão aplicados métodos de seleção de variáveis com o intuito de melhorar a interpretabilidade e eficiência dos modelos.

Análise Descritiva

Para uma análise inicial das variáveis explicativas, construímos boxplots comparando a distribuição de cada variável contínua em relação à variável resposta `Class`. A ideia é investigar visualmente se alguma variável apresenta padrões de separação claros entre as classes, o que pode indicar seu potencial preditivo.





Ao observar os gráficos, identificamos algumas variáveis com distribuições significativamente distintas entre as classes. Entre as variáveis `V1` a `V12`, destacam-se `V6`, `V7` e `V8` como aquelas que apresentam maior separação entre os grupos. Já no intervalo de `V13` a `V24`, observamos que `V14`, `V15` e as variáveis de `V17` a `V20` demonstram diferenças expressivas entre as classes. Essas variáveis são fortes candidatas a explicarem a variável resposta e devem ser consideradas com atenção nos modelos de classificação.

Metodologia

Como estratégia inicial, quatro modelos de regressão logística multinomial foram ajustados: (i) com todas as variáveis explicativas; (ii) com todas as variáveis e seus termos quadráticos; (iii) com as variáveis estatisticamente significativas do modelo completo; e (iv) com todas as variáveis, seus termos quadráticos e interações entre variáveis previamente identificadas como mais relevantes. A Tabela a seguir apresenta os valores de AIC, log-verossimilhança e acurácia para os quatro modelos.

```
## # A tibble: 4 x 4
##   Modelo                                AIC LogLik Acuracia
##   <chr>                               <dbl>  <dbl>    <dbl>
## 1 Modelo 1: Variáveis originais      7603. -3726.    0.712
## 2 Modelo 2: Termos quadráticos       3193. -1450.    0.910
## 3 Modelo 3: Variáveis significativas 7603. -3726.    0.712
## 4 Modelo 4: Quadráticos + Interações 3107. -1299.    0.916
```

Observa-se que o Modelo 4, que incorpora tanto termos quadráticos quanto interações entre variáveis selecionadas, apresentou o melhor desempenho entre todos os modelos testados, com menor AIC (3107.127),

maior log-verossimilhança (-1298.564) e acurácia de 91,6%. Isso indica que a adição controlada de complexidade ao modelo, por meio de interações entre variáveis relevantes, permitiu capturar melhor os padrões da base de dados, superando inclusive o modelo com apenas termos quadráticos. A análise prossegue com base nesse modelo mais robusto.

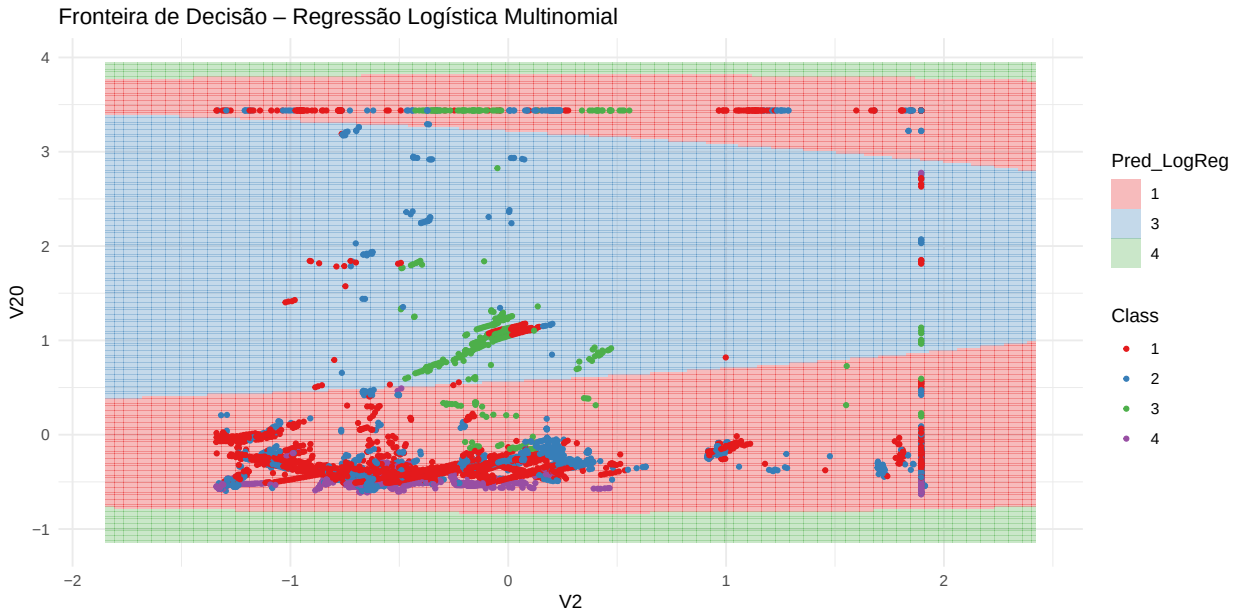
Além das abordagens baseadas em regressão logística multinomial, também foi testada uma técnica de redução de dimensionalidade baseada na Análise de Componentes Principais (PCA), aplicada sobre os dados padronizados. No entanto, apesar de simplificar o espaço de variáveis, a utilização do PCA resultou em desempenho inferior, tanto em termos de AIC quanto de acurácia, quando comparado aos modelos anteriores. Assim, optou-se por não prosseguir com essa abordagem. Como próximo passo, será ajustado o algoritmo k-Nearest Neighbors (KNN), visando avaliar seu desempenho em comparação com os modelos de regressão.

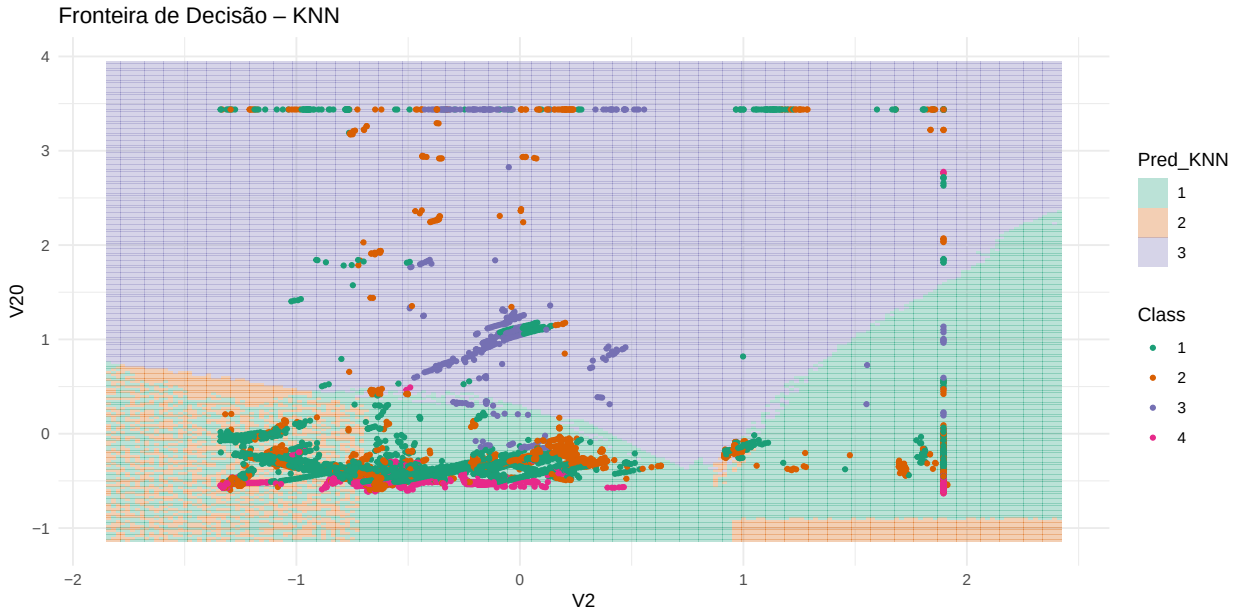
```
##     k  Accuracy     Kappa AccuracySD    KappaSD
## 1   5 0.8678538 0.8006728 0.01416649 0.02117884
## 2   7 0.8550212 0.7811090 0.01640948 0.02437212
## 3   9 0.8451176 0.7663212 0.01514379 0.02234622
## 4  11 0.8390699 0.7573039 0.01834147 0.02719359
## 5  13 0.8328432 0.7480843 0.01805081 0.02672770
## 6  15 0.8289926 0.7424262 0.01666826 0.02465362
## 7  17 0.8260612 0.7379738 0.01522057 0.02240829
## 8  19 0.8251454 0.7365370 0.01434118 0.02148341
## 9  21 0.8222140 0.7324162 0.01357335 0.02034855
## 10 23 0.8185480 0.7268014 0.01543454 0.02290951
```

```
## 5-nearest neighbor model
## Training set outcome distribution:
##
##    1    2    3    4
## 2205 2097  328  826
```

```
## # A tibble: 1 x 2
##   Modelo           Acuracia
##   <chr>               <dbl>
## 1 KNN (CV 10-fold)    0.868
```

O modelo KNN, ajustado com validação cruzada 10-fold, obteve uma acurácia de 86,8%, sendo otimizado com $k = 5$ vizinhos. Esse desempenho é ligeiramente inferior ao da regressão logística multinomial com termos quadráticos e interações (Modelo 4), que alcançou 91,6% de acurácia, mas ainda assim supera os modelos base com variáveis originais. Isso demonstra que o KNN é competitivo em tarefas de classificação multiclasse, especialmente quando bem ajustado, embora a regressão logística mais complexa tenha capturado melhor os padrões dos dados. Abaixo, temos a fronteira de decisão para cada uma das abordagens.





A imagem apresenta as fronteiras de decisão dos dois modelos de classificação aplicados: Regressão Logística Multinomial (acima) e KNN (abaixo), com base nas variáveis *V2* e *V20*. Observa-se que a fronteira gerada pela Regressão Logística é mais suave e tende a ser mais generalista, com regiões amplas associadas a cada classe. Esse comportamento é típico de modelos paramétricos, que assumem uma forma funcional fixa para separar as classes. Já o KNN exibe fronteiras mais irregulares e fragmentadas, adaptando-se fortemente à distribuição dos pontos de treino, o que sugere uma maior sensibilidade à variação local dos dados. No entanto, essa flexibilidade também pode tornar o modelo mais suscetível ao overfitting, especialmente em regiões com maior sobreposição entre classes.

Para garantir uma avaliação robusta dos modelos, foi adotado o método de validação cruzada estratificada do tipo k-fold com $k = 10$. Esse procedimento consiste em particionar os dados em 10 subconjuntos aproximadamente iguais, utilizando 9 partes para o treinamento e 1 para o teste, de forma rotativa. A estratificação assegura que a proporção entre as classes seja mantida em cada uma das dobras, tornando a comparação entre os modelos mais justa e confiável. Além da acurácia, foram calculadas outras métricas como o índice Kappa, a média do F1-score, a sensibilidade e a precisão, proporcionando uma visão mais

abrangente do desempenho de cada método.

```
## # A tibble: 2 x 6
##   Modelo                          Kappa Accuracy F1_Média Sensibilidade Precisao
##   <chr>                           <dbl>    <dbl>    <dbl>         <dbl>    <dbl>
## 1 Regressão Multinomial - Modelo~ 0.858    0.906    0.901         0.896    0.907
## 2 KNN (melhor k, CV 10-fold)      0.797    0.866    0.860         0.858    0.863
```

Conclui-se, portanto, que a Regressão Logística Multinomial com termos quadráticos e interações (Modelo 4) apresentou o melhor desempenho geral, tanto em termos de acurácia (90,6%) quanto de estabilidade na classificação, como evidenciado pelas métricas obtidas na validação cruzada. Sua principal vantagem está na capacidade de modelar relações não lineares e interações entre variáveis de forma interpretável e controlada. Por outro lado, sua limitação reside na suposição de linearidade (mesmo com extensões quadráticas) e na necessidade de especificar previamente essas relações. Já o KNN, apesar de mais simples e intuitivo, depende fortemente da densidade dos dados em cada região do espaço, sendo menos robusto em cenários com alta dimensionalidade ou sobreposição de classes.

Referências

Cover, T. M., & Hart, P. E. (1967). *Nearest neighbor pattern classification.* IEEE Transactions on Information Theory, 13(1), 21–27. https://doi.org/10.1109/TIT.1967.1053964

Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied Logistic Regression* (3rd ed.). Wiley. ISBN: 978-0470582473.