Métodos de machine learning no data set Covertype

Gonçalo Rocha nº202000086

Bioinformática

Aprendizagem Automática

Descrição dos dados

O conjunto de dados concentra-se na previsão de tipos de cobertura florestal em quatro áreas selvagens na Floresta Nacional de Roosevelt, Colorado.

Inclui variáveis cartográficas e colunas binárias para variáveis independentes qualitativas, como áreas selvagens e tipos de solo.

Descrição dos dados

- Tem 581,012 linhas

- Tem 12 medidas, mas 54 colunas
  - 10 variáveis quantitativas
  - 4 binarias sobre as áreas selvagens
  - 40 binárias sobre tipos de solo

- Não tem missings

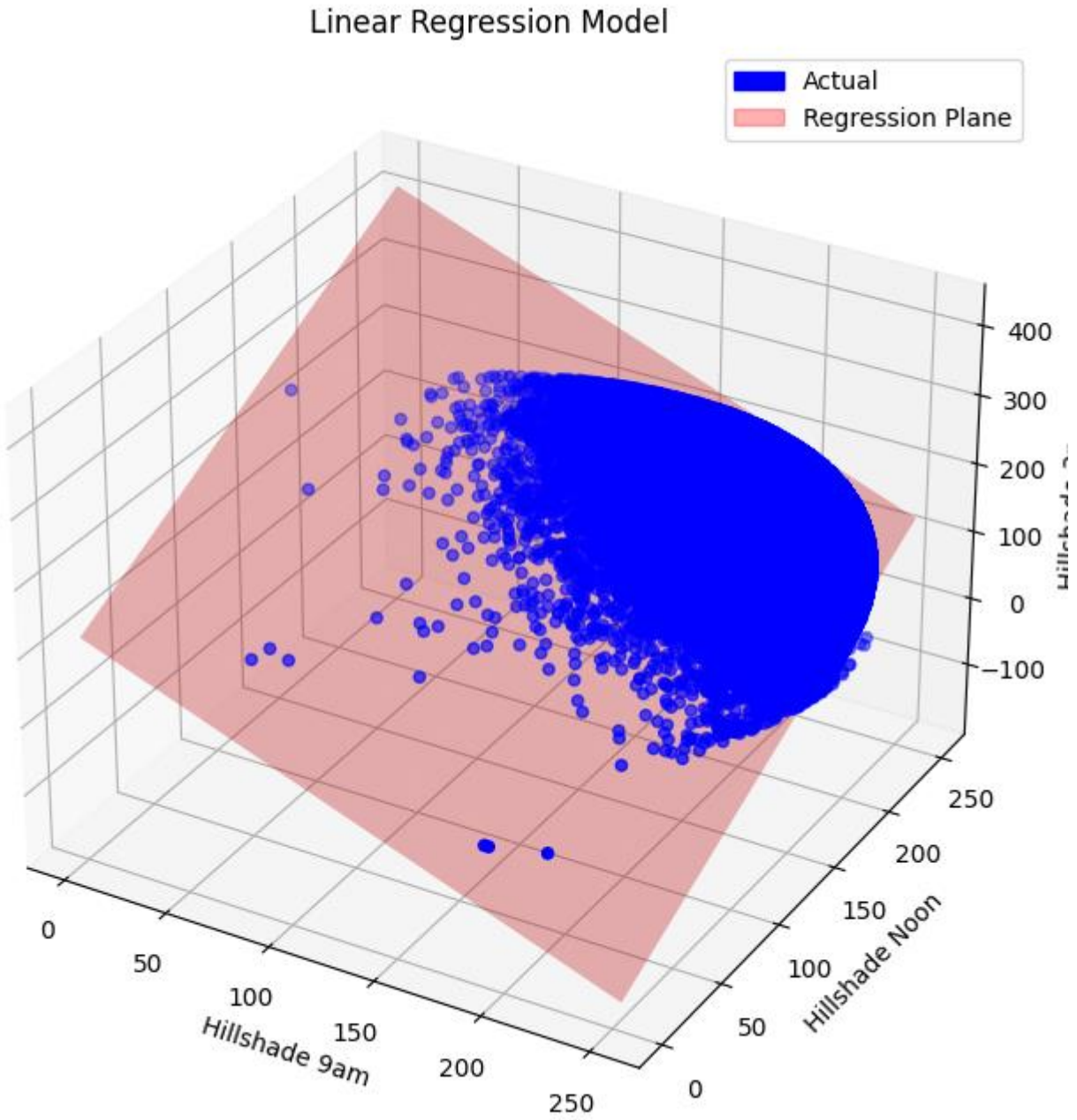
Descrição dos dados

- As colunas (variáveis) consideraras são:
  - Elevação (em metros)
  - Exposição (em graus azimute)
  - Declive (em graus)
  - Distância horizontal para hidrologia (em metros)
  - Distância vertical para hidrologia (em metros)
  - Distância vertical para estradas (em metros)
  - Sombreamento 09:00h (em graus)
  - Sombreamento 12:00h (em graus)
  - Sombreamento 15:00h (em graus)
  - Distancia horizontal para pontos de incendio (em metros)

Modelo de Regressão Linear

- Sombreamento 09:00h
- Sombreamento 12:00h
- Targert - Sombreamento 15:00h

Regressão Linear múltipla



Regressão Linear Multipla

Erro Quadrático Médio(MSE)= 43.0689

A diferença entre os valores reais os valores previstos

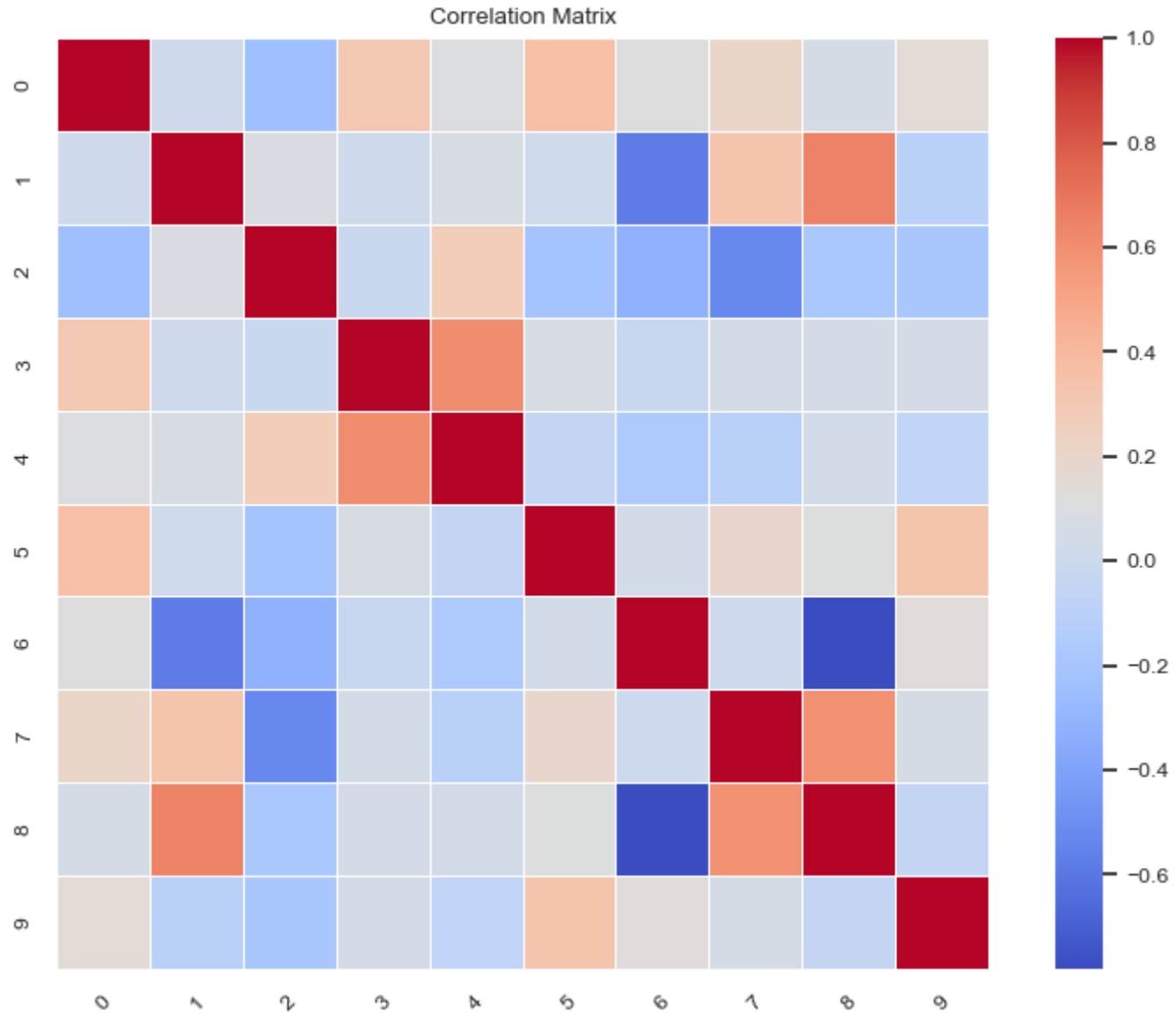
Coeficiente de determinação (R2) = 0.9708

97,08% da variância na variável dependente é explicada pelas variáveis independentes

Regressão linear Multipla

O baixo MSE e o alto valor de R-quadrado sugerem que o modelo é capaz de fazer previsões precisas e capturar uma quantidade significativa da variação na variável dependente com base nas duas variáveis independentes usadas.

Matriz de Correlação



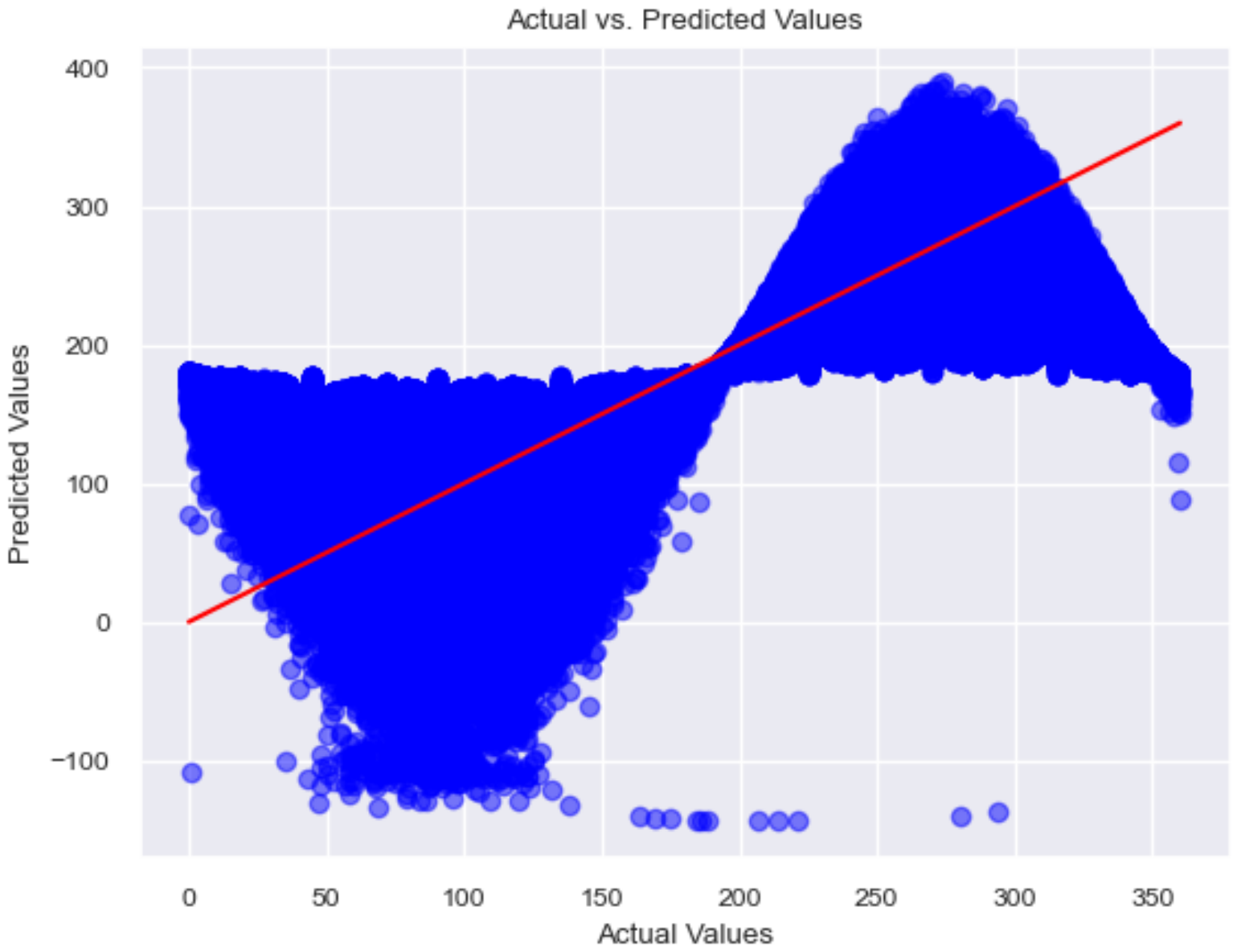
Random Forest Regression

Para a Random Forest Regression foi usado a feature 7 e 8, que correspondem a:

- Hillshade 12:00h
- Hillshade 15:00h

E at target feature 1 corresponde a:
- Aspect

Random Forest Regression



Random Forest Regression

Raiz do erro quadrático médio(RMSE) = 85,18

Erro Absoluto Médio (MAE) = 66,90

R-quadrado (R^2) = 0,05

Pode-se concluir que o modelo pode não estar a ter um bom desempenho em prever com precisão a variável alvo contínua

Fim