

Avaliação de técnicas probabilísticas para regressão em dados de características pessoais e fatores dietéticos





Aluno: Gletson Girão

ROTEIRO

- Contextualização dos problemas
- Descrição dos dados
- Detalhamento da metodologia
- Aplicação dos modelos
 1. Exploração
 2. Preparação
 3. Aplicação dos modelos
- Avaliação dos resultados

Contextualização do Problema

active  ARFF  Publicly available  Visibility: public  Uploaded 29-09-2014 by [Joaquin Vanschoren](#)

 0 likes  downloaded by 2 people, 2 total downloads  0 issues  0 downvotes

 **OpenML-Reg19** **study_239**  [Add tag](#)

 Loading wiki

Author:

Source: Unknown - Date unknown

Please cite:

Determinants of Plasma Retinol and Beta-Carotene Levels

Summary:

Observational studies have suggested that low dietary intake or low plasma concentrations of retinol, beta-carotene, or other carotenoids might be associated with increased risk of developing certain types of cancer. However, relatively few studies have investigated the determinants of plasma concentrations of these micronutrients. We designed a cross-sectional study to investigate the relationship between personal characteristics and dietary factors, and plasma concentrations of retinol, beta-carotene and other carotenoids. Study subjects (N = 315) were patients who had an elective surgical procedure during a three-year period to biopsy or remove a lesion of the lung, colon, breast, skin, ovary or uterus that was found to be non-cancerous. We display the data for only two of the analytes.

Plasma concentrations of the micronutrients varied widely from subject to subject. While plasma retinol levels varied by age and sex, the only dietary predictor was alcohol consumption ($R^2 = .38$). Plasma beta-carotene levels were log-transformed prior to the analyses due to severe asymmetry of the residuals on the original scale. For log beta-carotene, dietary intake, regular use of vitamins, and intake of fiber were associated with higher plasma concentrations, while Quetelet Index (defined as $\text{weight}/\text{height}^2$ in the units kg/m^2) and cholesterol intake were associated with lower plasma levels, adjusting for the other factors ($R^2 = .50$). There was one extremely high leverage point in alcohol consumption that was deleted prior to the analyses. Plasma concentrations of retinol and beta-carotene were not correlated.

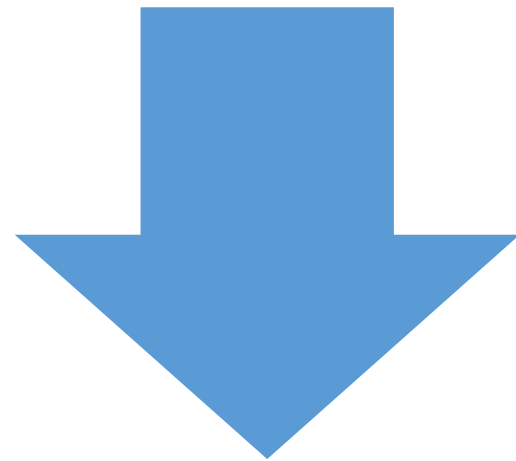
We conclude that there is wide variability in plasma concentrations of these micronutrients in humans, and that much of this variability is associated with dietary habits and personal characteristics. A better understanding of the physiological relationship between some personal characteristics and plasma concentrations of these micronutrients will require further study.

Authorization: Contact Authors

Reference: These data have not been published yet but a related reference is

Nierenberg DW, Stukel TA, Baron JA, Dain BJ, Greenberg ER. Determinants of plasma levels of beta-carotene and retinol. American Journal of Epidemiology 1989;130:511-521.

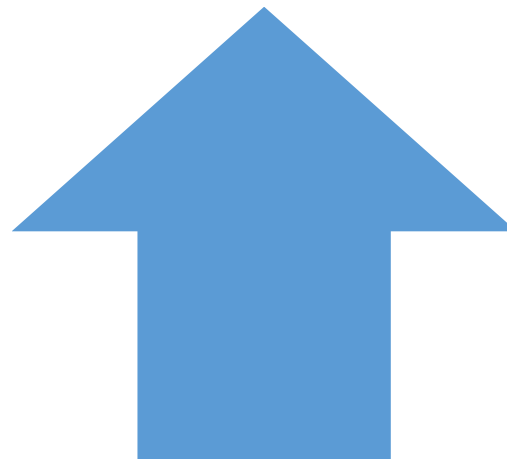
Contextualização do problema



- Retinol
- Beta-Caroteno
- Outros Carotenoides



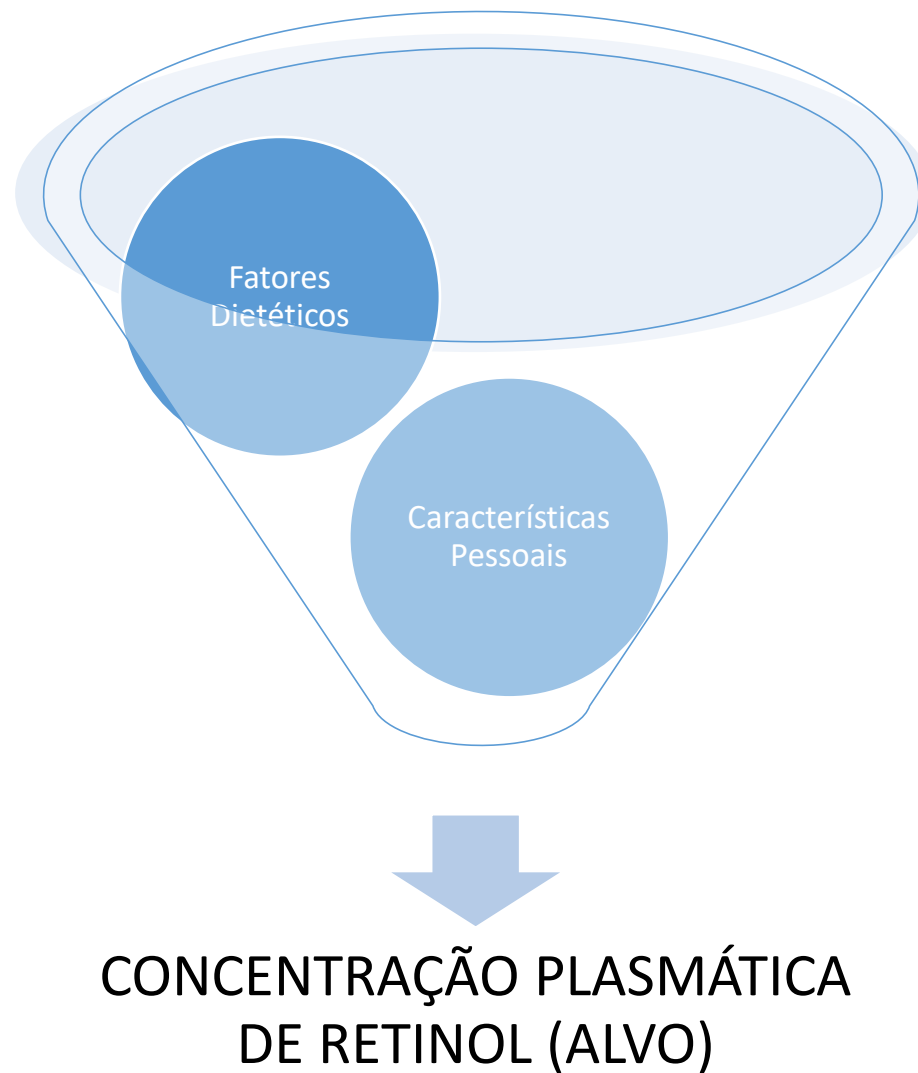
Risco de
desenvolver
Câncer



Estudos observacionais sugerem que baixas concentrações plasmáticas de retinol, betacaroteno ou outros carotenoides podem estar associados a um maior risco de desenvolver câncer.

Contextualização do problema

O problema abordado consiste em descobrir a relação entre características pessoais e fatores dietéticos (preditores) e a concentração plasmática de retinol (alvo).



Descrição dos dados

| | ATRIBUTOS | DESCRIÇÃO |
|-------------|-------------|---|
| CATEGÓRICOS | VITUSE | Uso de vitamina (1=Sim, com frequência, 2=Yes, pouca frequência, 3=Não) |
| | SEX | Sexo (1=Masc, 2=Fem) |
| | SMOKSTAT | Status de fumante (1=Nunca, 2=Ex-fumante, 3=Fumante) |
| | AGE | Idade em anos |
| | ALCOHOL | Quantidade de drinks alcoólicos consumidos por semana |
| NUMÉRICOS | CALORIES | Calorias consumidas por dia |
| | FAT | Gramas de gordura consumidas por dia |
| | FIBER | Gramas de fibra consumidas por dia |
| | QUETELET | IMC (PESO/(ALTURA^2)) |
| | CHOLESTEROL | Colesterol consumido (mg por dia) |
| | BETADIET | Betacaroteno na dieta consumido (mcg por dia) |
| ALVO | RETDIET | Retinol na dieta consumido (mcg por dia) |
| | BETAPLASMA | Plasma Betacaroteno (ng/ml) |
| | RETPLASMA | Plasma Retinol (ng/ml) |

DETALHAMENTO DA METODOLOGIA

PRIMEIRA PARTE

Avaliação do modelo de regressão linear bayesiana (Implementação própria), comparado ao modelo de regressão linear convencional do scikit-learn

SEGUNDA PARTE

Uso de um modelo de árvore mais complexo (Floresta Aleatória) para avaliar técnica de Otimização Bayesiana.

Ferramentas

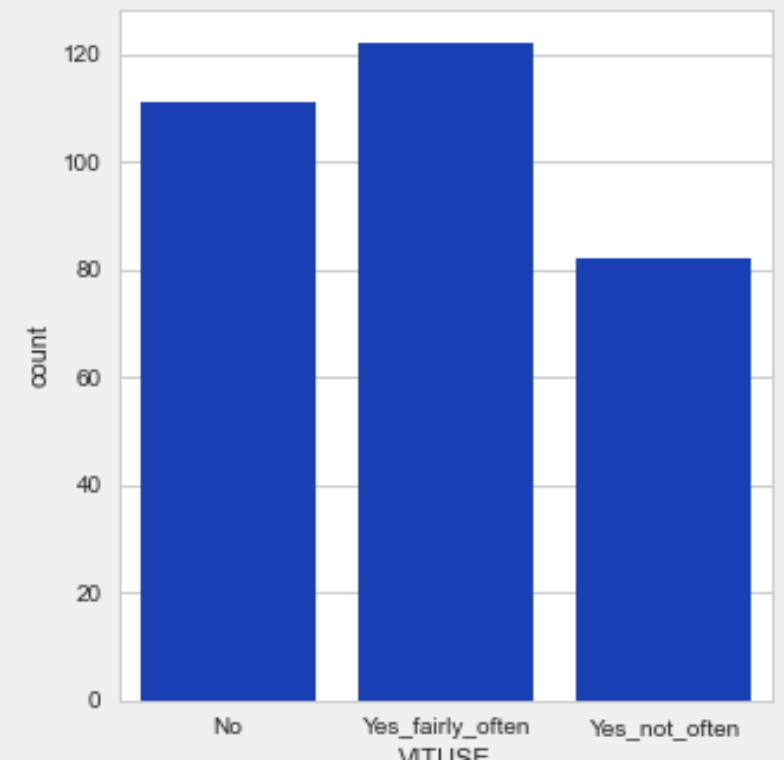
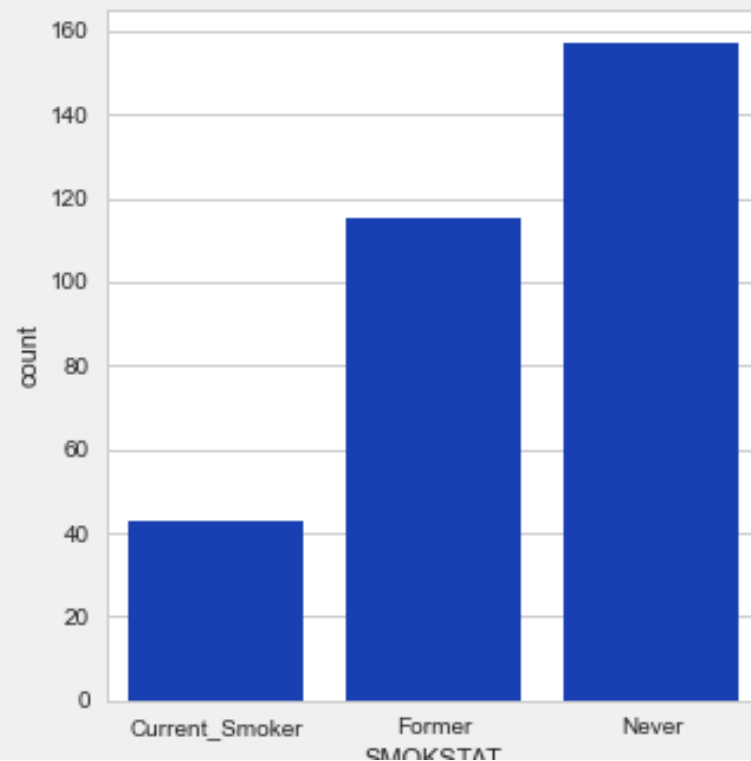
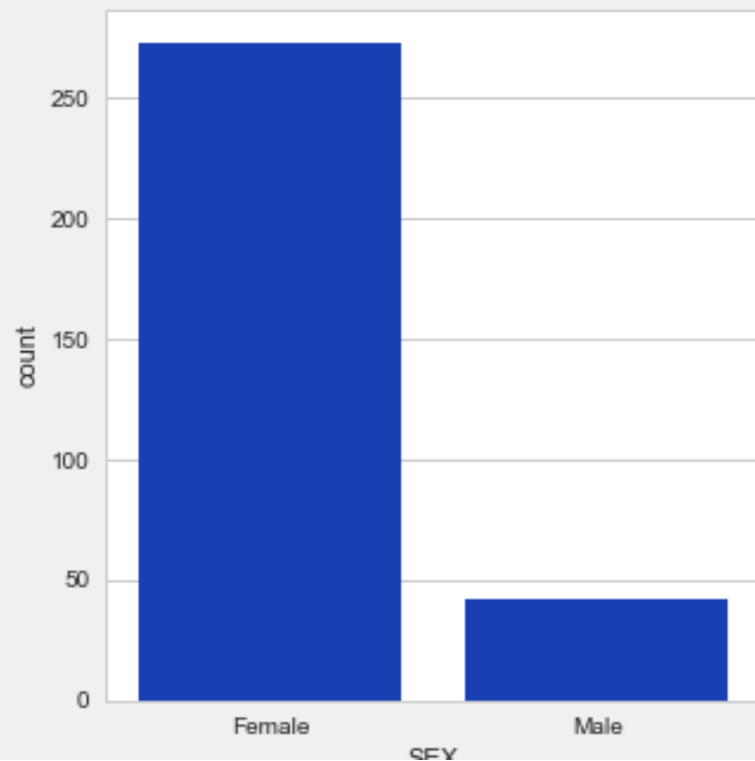
Para as tarefas de tratamento, exploração, pré-processamento, seleção e avaliação de modelos: bibliotecas pandas e scikit-learn;

Biblioteca bayes-opt para otimização.

Aplicação dos Modelos

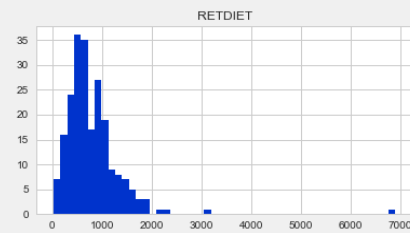
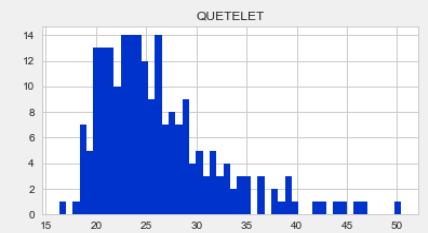
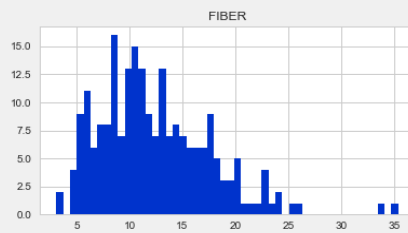
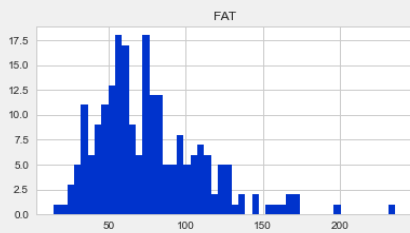
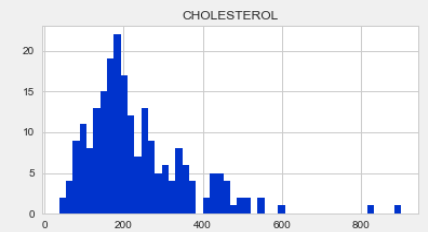
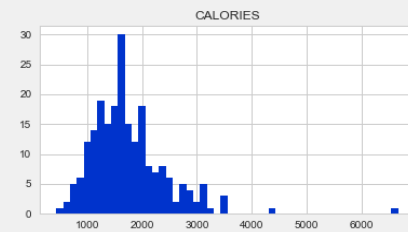
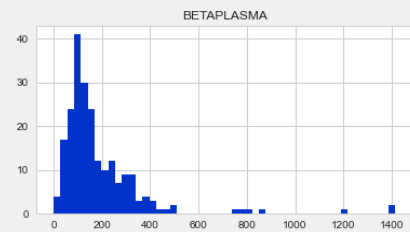
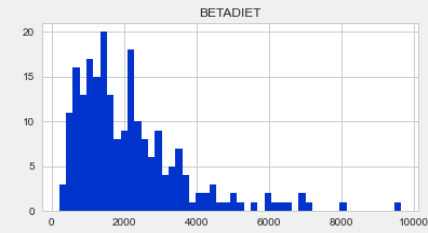
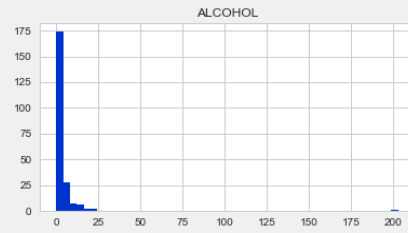
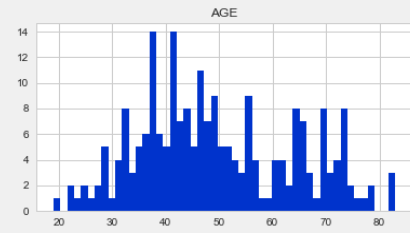
Exploração dos dados

Contagem dos atributos categóricos

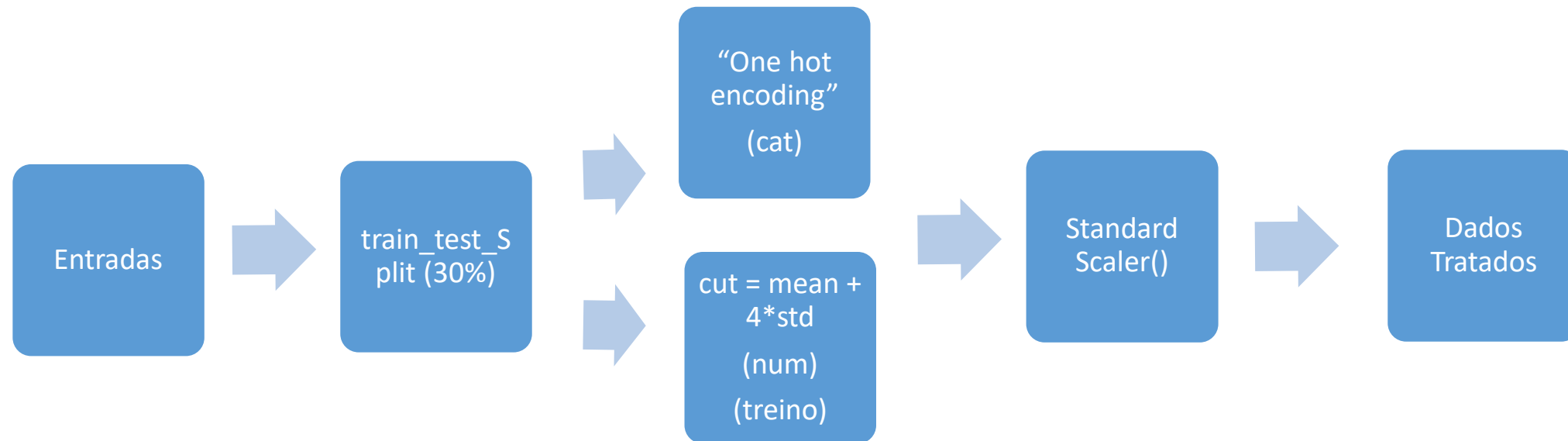


Exploração dos dados

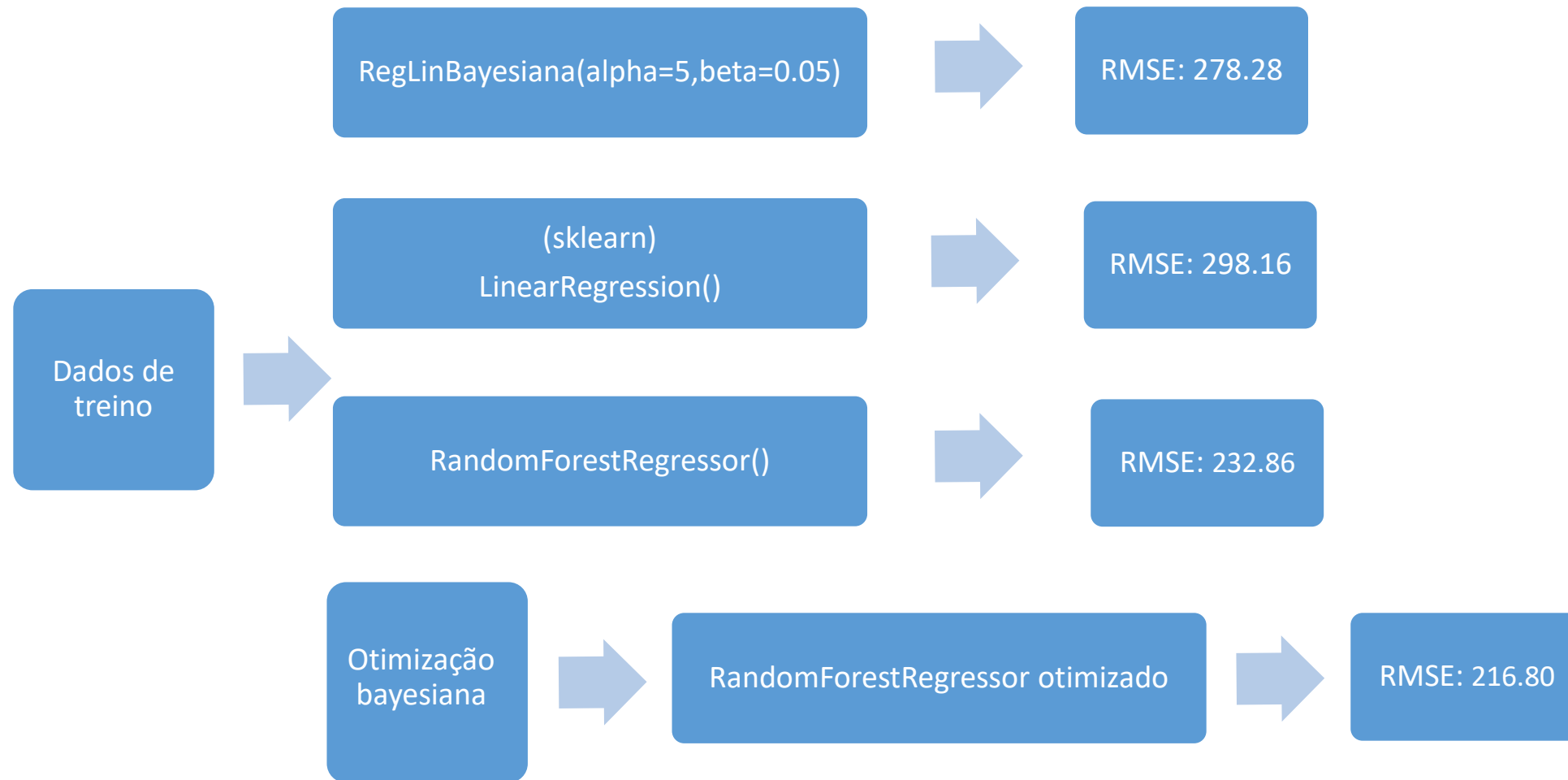
Observando a distribuição dos atributos numéricos



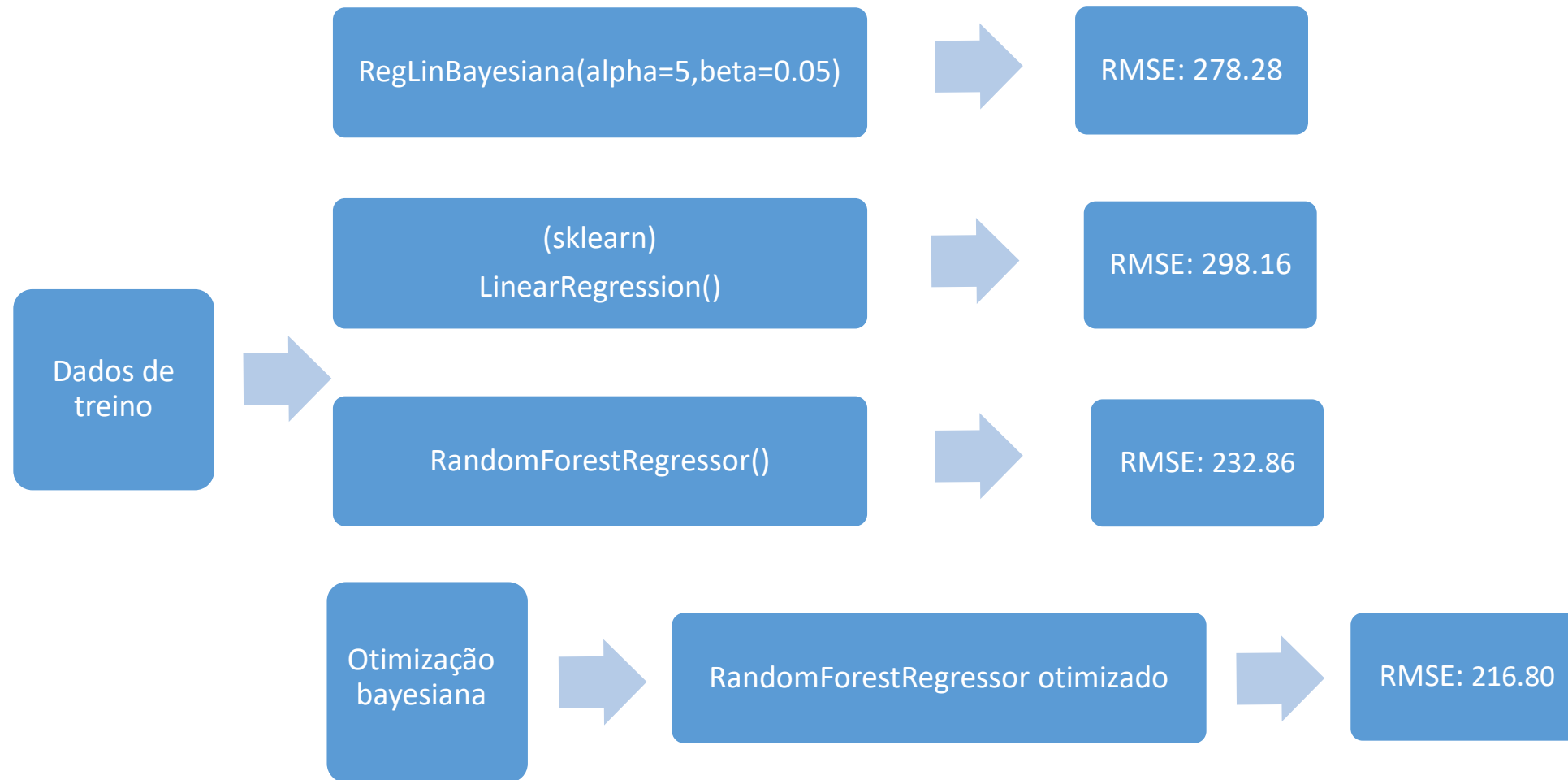
Preparação dos dados



Aplicação dos modelos



Aplicação dos modelos



Avaliação dos Resultados

PRIMEIRA PARTE

O modelo bayesiano tem rmse 278.28 e o modelo convencional tem rmse 298.16

A regressão bayesiana tem uma performance levemente melhor que o modelo convencional para este caso.

Além de melhorar a métrica de avaliação o modelo probabilístico ainda retorna uma medida de incerteza para cada valor predito.

A aplicação do modelo com atributos polinomiais teve um pior desempenho para os dois modelos.

Avaliação dos Resultados

SEGUNDA PARTE

O modelo mais complexo de árvores tem uma performance consideravelmente melhor para a tarefa que os modelos anteriores.

O modelo otimizado tem uma performance levemente melhor, rmse 216.80, face ao modelo com parâmetros padrão da biblioteca com rmse 232.86.

CONSIDERAÇÕES FINAIS

- Aplicação de um modelo de regressão probabilístico que mesmo com uma implementação simples atingiu um bom resultado perante ao equivalente convencional.
- Aplicação com sucesso de procedimento de otimização bayesiana para otimização de hiperparâmetros de modelos de aprendizado de máquina, procedimento esse que é parte importante do processo de aprendizado de máquina e pode ser utilizado em diversos outros contextos.
- Muitas melhorias podem ser feitas: engenharia de atributos, seleção de variáveis, correlações.

OBRIGADO!