Universidade Federal do Rio de Janeiro

Projeto do Curso
Disponível em:
https://github.com/Lemos-san/-COE24-_estatistica_modelos_probabilisticos

Professor: Rosa Maria Leão Período: 2018/2
Disciplina: COE241 - Estatística e Modelos Probabilísticos

Matheus Lemos dos Reis DRE: 114086213

1. Introdução

1.1. Propósito do Documento

Este documento representa o trabalho de conclusão da disciplina 'Estatística e Modelos Probabilísticos' [COE241], no período 2018/2. O objetivo deste documento é estudar um conjunto de dados aplicando a teoria aprendida em classe.

2. Descrição Geral

2.1. Perspectiva do Projeto

Foram analisados dados reais fornecidos pelo Professor Clauido Gil Soares de Araujo da CLINIMEX, através da aluna de doutorado da UFRJ Christina G. de Souza e Silva. Os dados foram obtidos a partir de uma extensa base de dados do Prof. Claudio Gil, coletada durante muitos anos e usada em suas pesquisas. Os dados mostram uma medida da condição aeróbica do paciente (o VO2 max) (por quilo de peso do indivíduo) e ainda as variáveis idade, peso e a carga máxima atingida durante um teste ao qual o paciente foi submetido. As características dos dados fornecidos são: idade do paciente, peso (kg), carga nal (watts) e VO2 máximo (mg/Kg/min). Todos os dados estão contidos no arquivo Dados-medicos.csv.

2.2. Ferramentas Utilizadas

Foi escolhido Python como linguagem de programação do projeto devido à grande quantidade de bibliotecas implementadas voltadas para o campo de estatística. Foi também utilizado o ambiente Jupyter como base principal para a execução do programa e a geração dos gráficos. Dentre as bibliotecas utilizadas, pode-se destacar:

- Numpy: cálculos numéricos e manipulações de matrizes
- Matplotlib: geração e visualização de gráficos
- Scipy: utilização de modelos de distribuições estatísticas
- Pandas: manipulação de dados

3. Análises

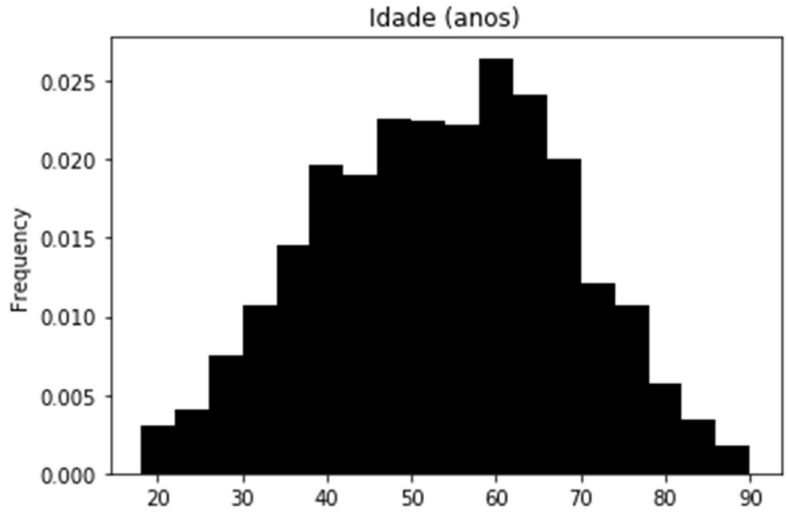
3.1. Histograma e Função Distribuição Empírica

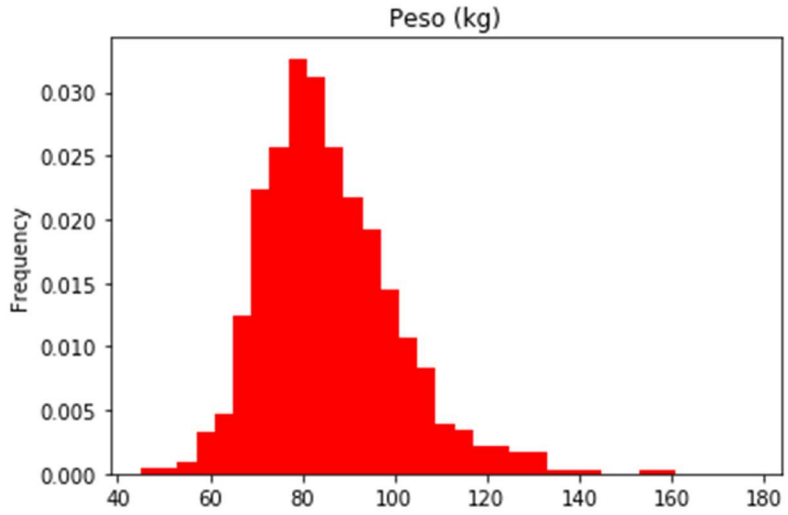
Pode-se notar nos histogramas como todas as distribuições das quatro variáveis parecem se aproximar de uma distribuição gaussiana. A variável de Idade, mais especificamente, parece ser a mais próxima deste comportamento.

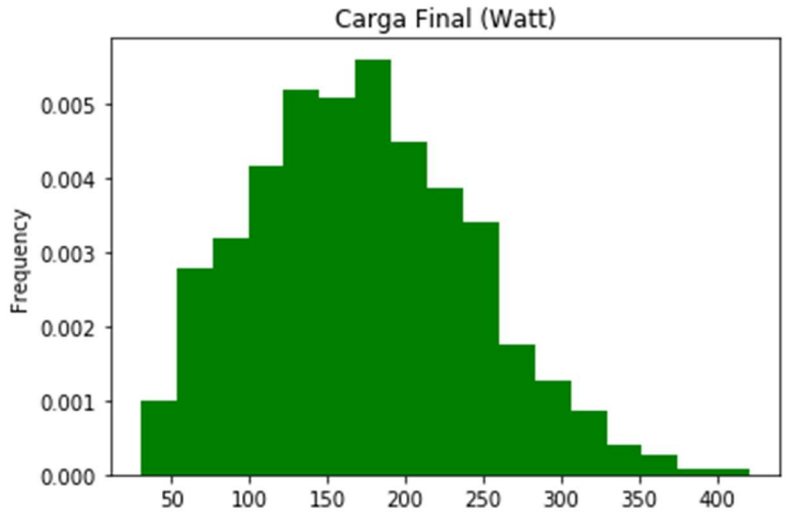
Os tamanhos e quantidades de *bins* de cada histograma foram determinados baseados nas fórmulas: (onde n é a quantidade de registros no conjunto de dados)

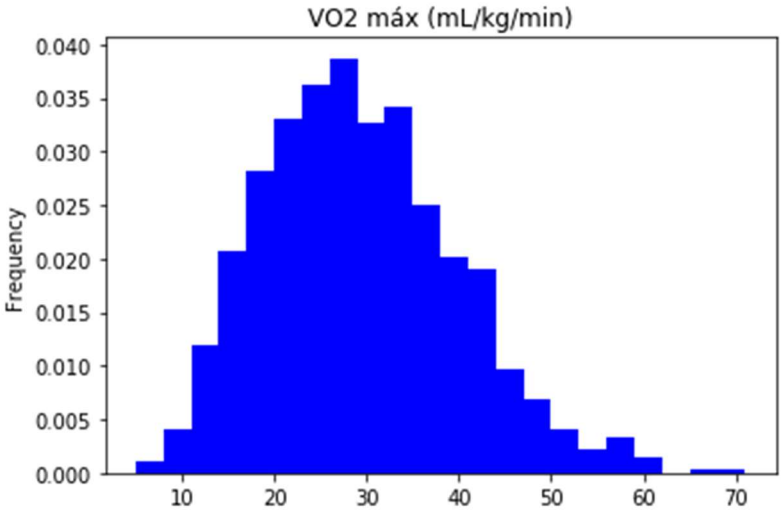
$$bins = (1 + 3.3 \times log_{10} n)$$

$$bin\ size = 3.49 \times \sigma \times n^{-1/3}$$

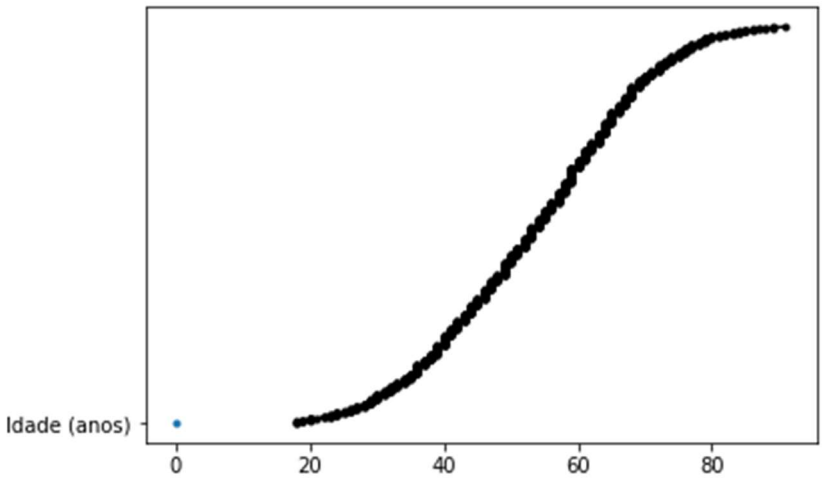


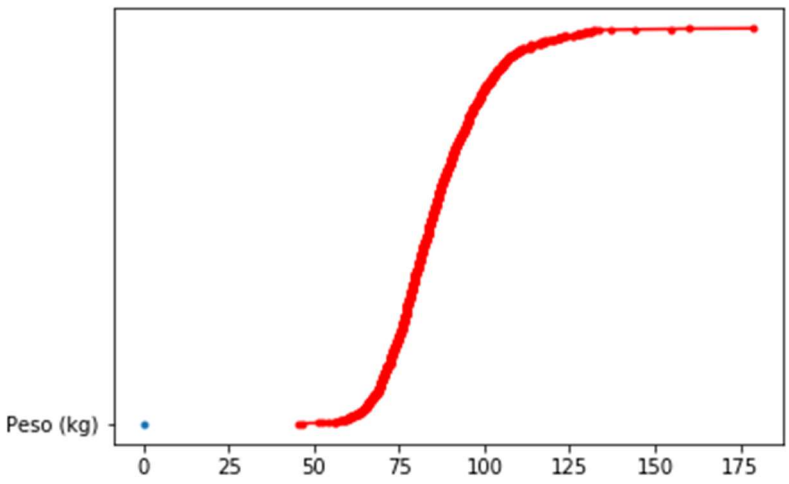


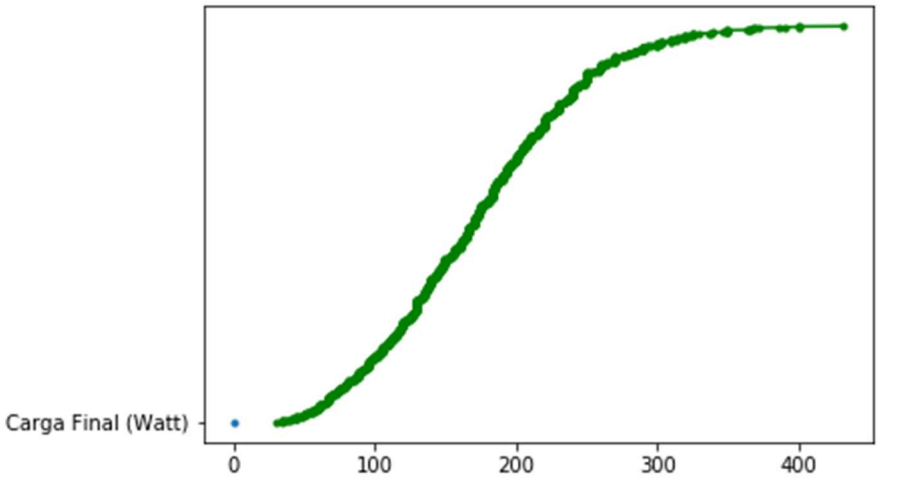


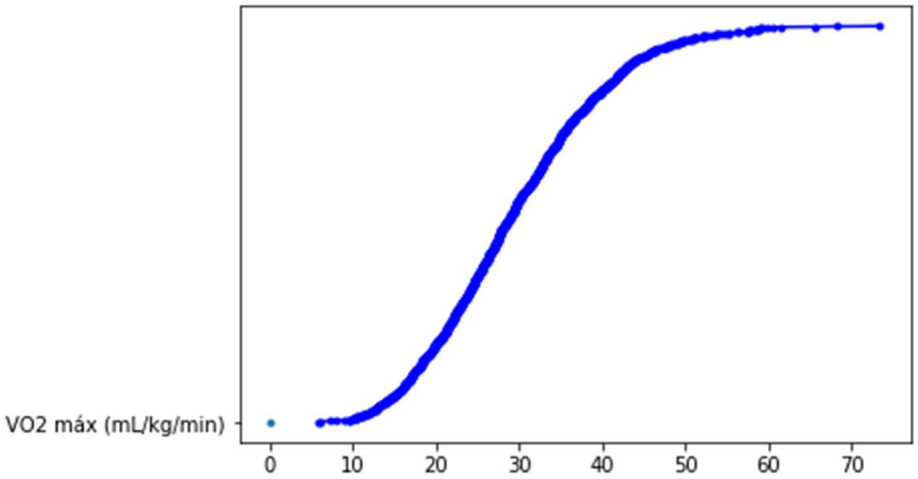


Abaixo seguem as respectivas Funções Distribuições Empíricas (CDFs) de cada variável. Pode-se notar como a curva da variável peso é muito mais “íngreme” do que as demais, o que condiz com sua distribuição mais concentrada no histograma. (o ponto azul claro no canto inferior esquerdo dos gráficos é apenas um ponto de referência para a construção dos gráficos, não fazendo parte de qualquer variável dos dados)







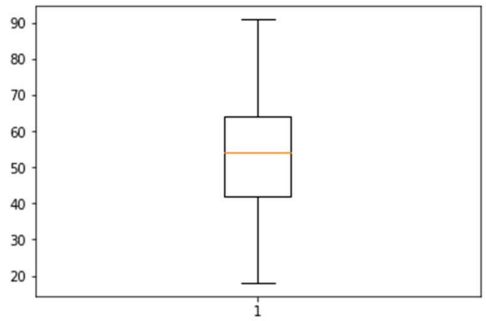


3.2. Média, Variância e Boxplot

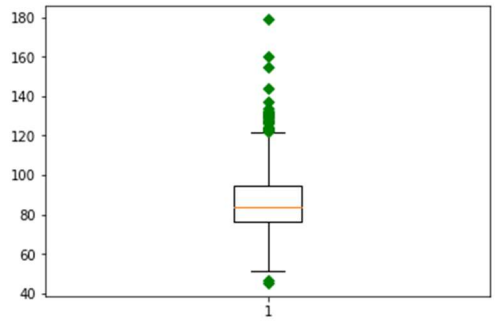
Variável	Média	Variância	Desvio Padrão
Idade	53.29095	217.26773	14.74000
Peso	85.92577	218.82688	14.79279
Carga final	172.27150	4908.85396	70.06321
VO2 máximo	29.39472	110.09823	10.49277

Pode-se notar como a média e variância de Carga final é muito mais alta que as das de Idade, Peso e VO2 máx. Nos gráficos BoxPlot abaixo, pode-se ver como as variáveis Peso e VO2 máx possuem a maior quantidade de *outliers* que poderia ser removidos para futuras análises a fim de não comprometer possíveis conclusões a serem tiradas do conjunto de dados em questão.

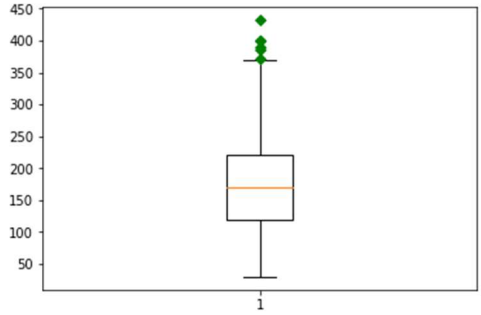
IDADE:



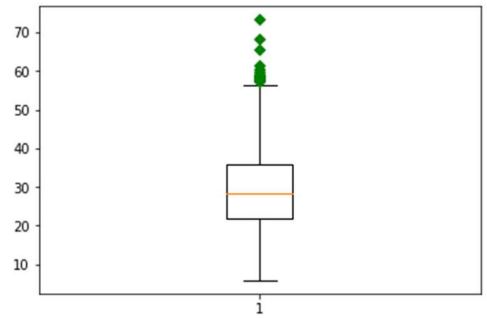
PESO:



CARGA FINAL:



VO2 MÁX:



3.3. Parametrizando distribuições

Foi utilizado o método da máxima verossimilhança para estimar os parâmetros das distribuições *gaussiana, exponencial, lognormal* e *weibull*. O método da máxima verossimilhança (MLE) supõe uma distribuição X contendo k parâmetros $\theta_1, \theta_2, \dots, \theta_k$ e uma pdf $f(x|\theta_1, \theta_2, \dots, \theta_k)$. Se possuirmos uma amostra aleatória $X_1, X_2, \dots, X_n$com valores observados como sendo $x_1, x_2, \dots, x_n$a cdf de $X_1, X_2, \dots, X_n$ é $\prod_{i=1}^{n} f(x_i|\theta_1, \theta_2, \dots, \theta_k)$.

Define-se a função de verossimilhança (*likelihood*) como sendo $L(\theta_1, \theta_2, \dots, \theta_k|x_1, x_2, \dots, x_k) = \prod_{i=1}^{n} f(x_i|\theta_1, \theta_2, \dots, \theta_k)$. Os valores dentre $\theta_1, \theta_2, \dots, \theta_k$ que maximizam a função, são chamados de estimadores máximos (MLE) dos parâmetros. Os valores MLE são aqueles para qual a sequência de amostras tem a maior probabilidade de ocorrer.

Segue abaixo, respectivamente, as MLE e as estimativas de cálculo dos parâmetros para as distribuições utilizadas:

Gaussiana/Normal:

$$f(x_1, \ldots, x_n \mid \mu, \sigma^2) = \prod_{i=1}^{n} f(x_i \mid \mu, \sigma^2) = \left(\frac{1}{2\pi\sigma^2}\right)^{n/2} \exp\left(-\frac{\sum_{i=1}^{n}(x_i - \mu)^2}{2\sigma^2}\right)$$

$$\hat{\mu} = \bar{x} = \sum_{i=1}^{n} \frac{x_i}{n}$$

Exponencial:

$$L(\lambda) = \prod_{i=1}^{n} \lambda \exp(-\lambda x_i) = \lambda^n \exp\left(-\lambda \sum_{i=1}^{n} x_i\right) = \lambda^n \exp(-\lambda n \overline{x})$$

$$\widehat{\lambda} = \frac{1}{\overline{x}} = \frac{n}{\sum_i x_i}$$

Lognormal:
(φ é função densidade da distribuição normal)

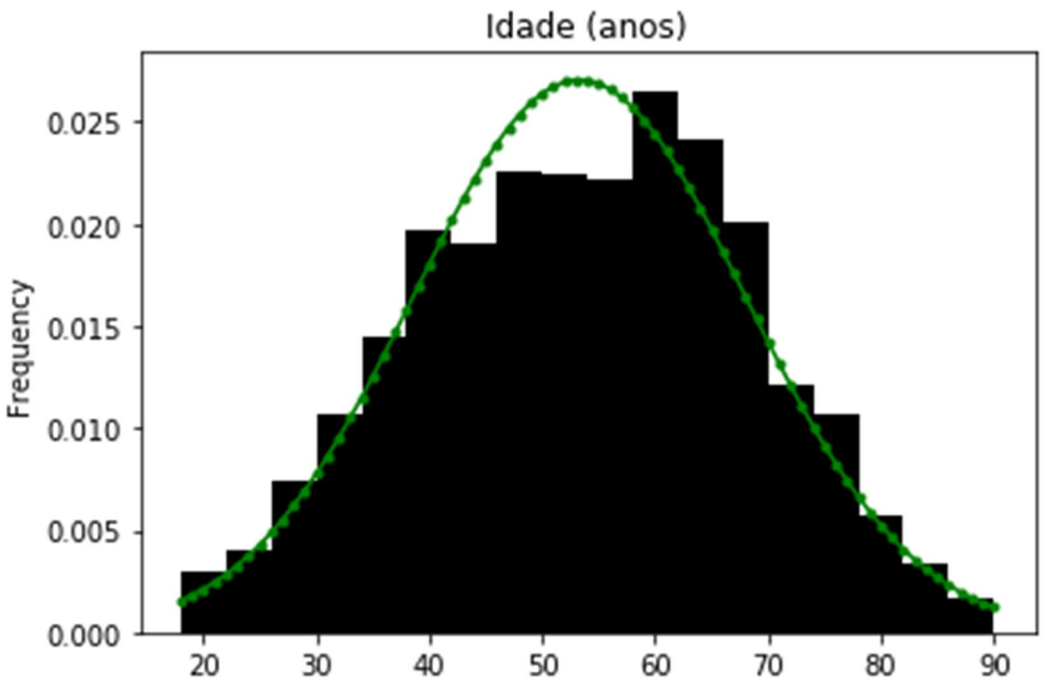
$$L(\mu, \sigma) = \prod_{i=1}^{n} \frac{1}{x_i} \varphi_{\mu,\sigma}(\ln x_i)$$

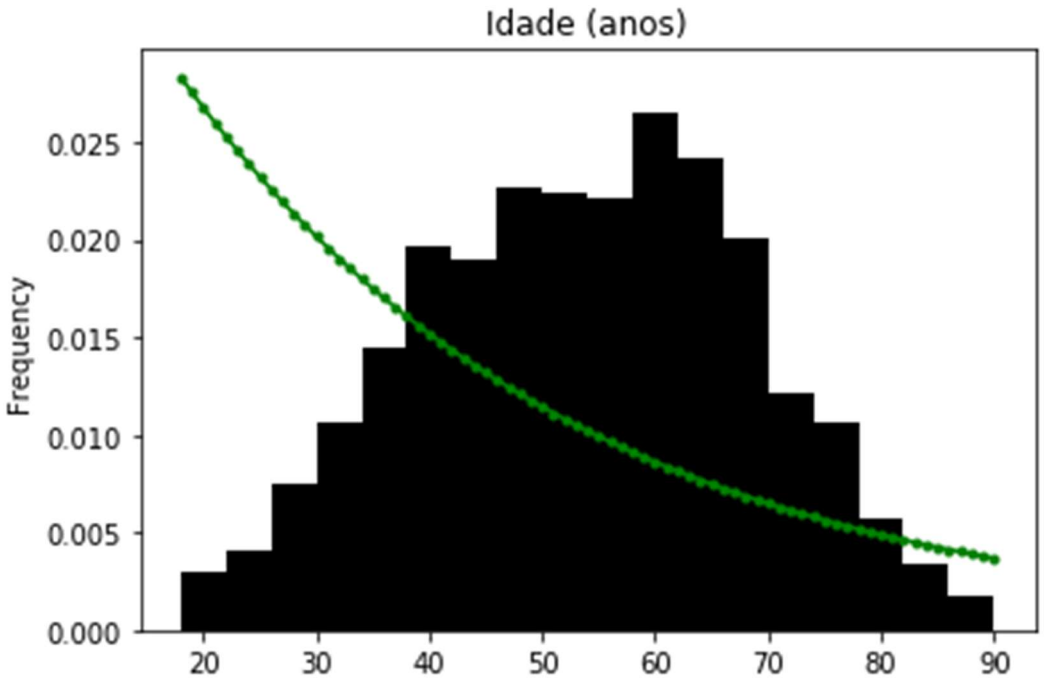
$$\widehat{\mu} = \frac{\sum_k \ln x_k}{n}, \qquad \widehat{\sigma}^2 = \frac{\sum_k (\ln x_k - \widehat{\mu})^2}{n}$$

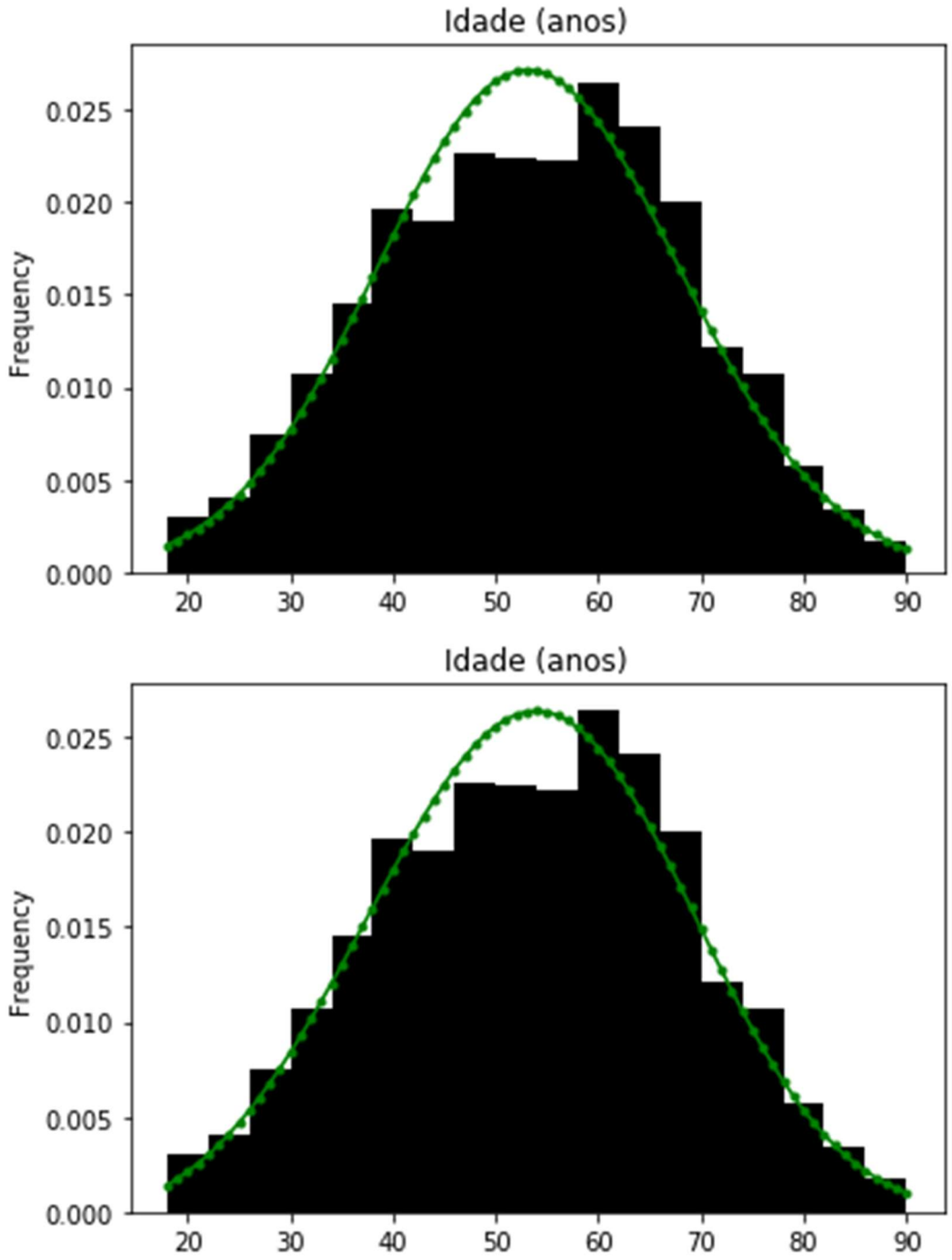
Weibull: (somente parâmetros)

$$\widehat{\lambda}^k = \frac{1}{n} \sum_{i=1}^{n} x_i^k$$

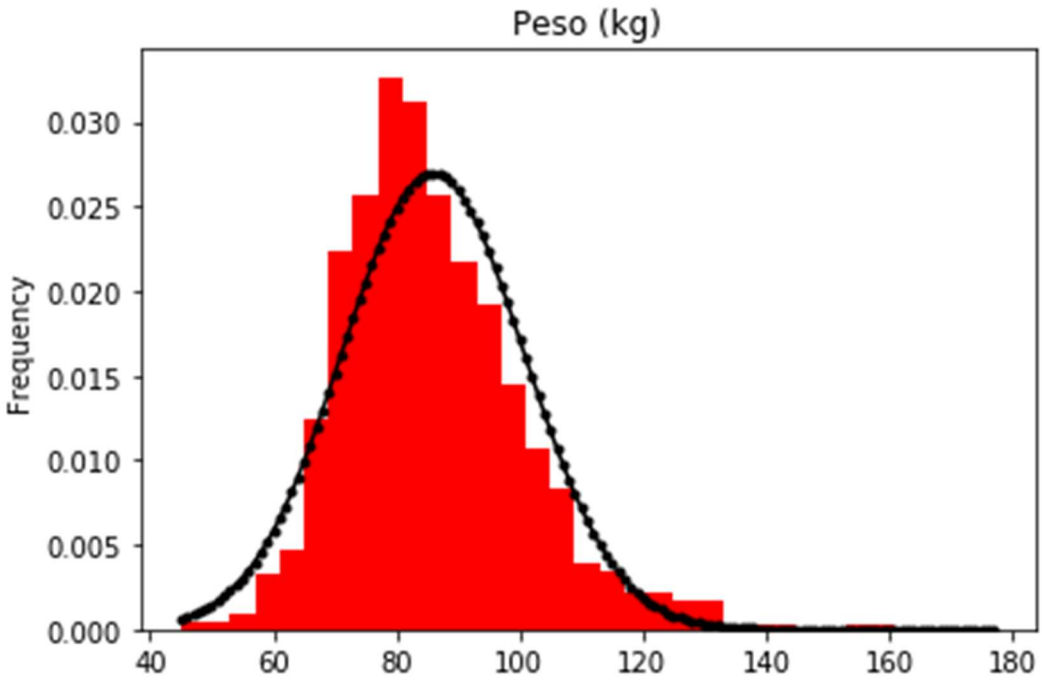
Os cálculos de dos parâmetrosSegue abaixo os gráficos dos histogramas sobrepostos pelas distribuições:
Distribuições em ordem: gaussiana, exponencial, lognormal, weibull.

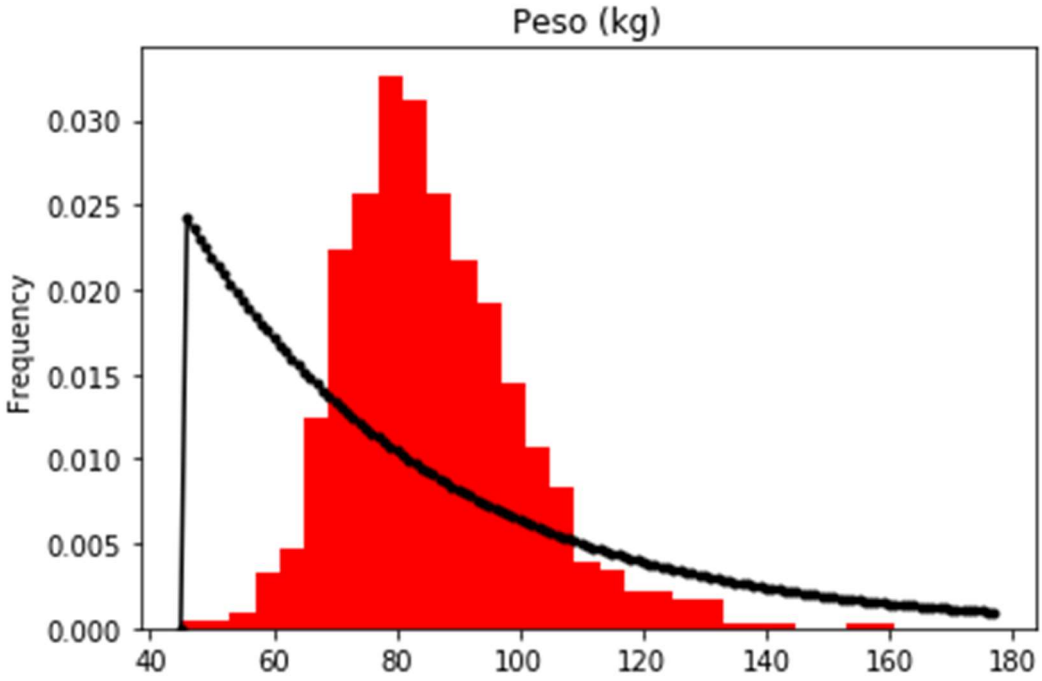

Idade (anos)
Frequency
0.025
0.020
0.015
0.010
0.005
0.000
20
30
40
50
60
70
80
90

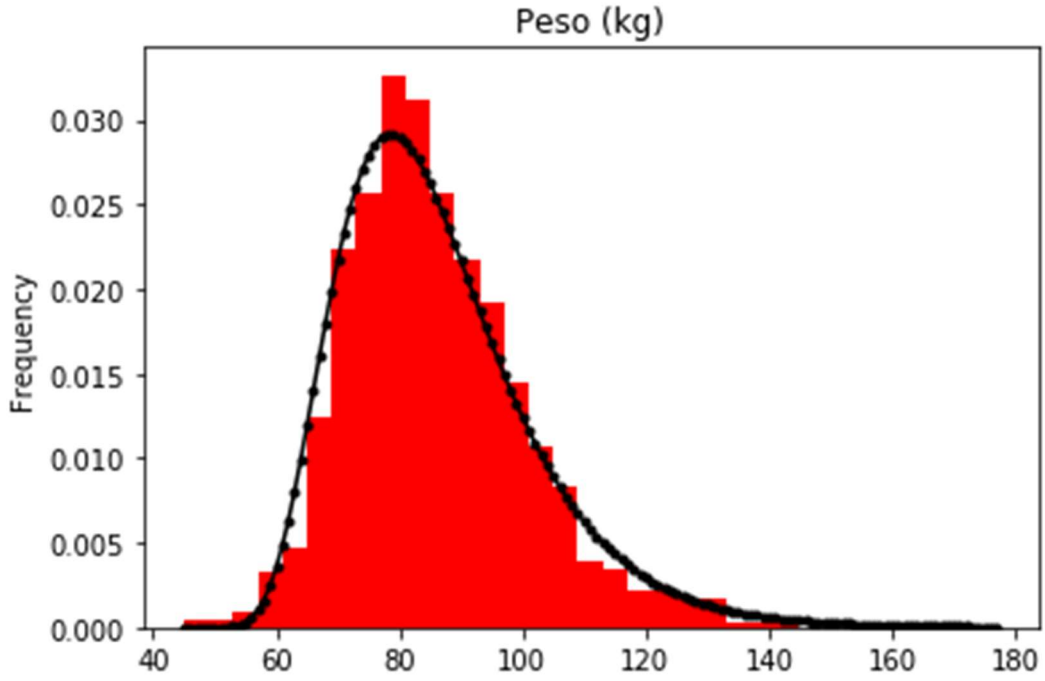

Idade (anos)
Frequency
0.025
0.020
0.015
0.010
0.005
0.000
20
30
40
50
60
70
80
90

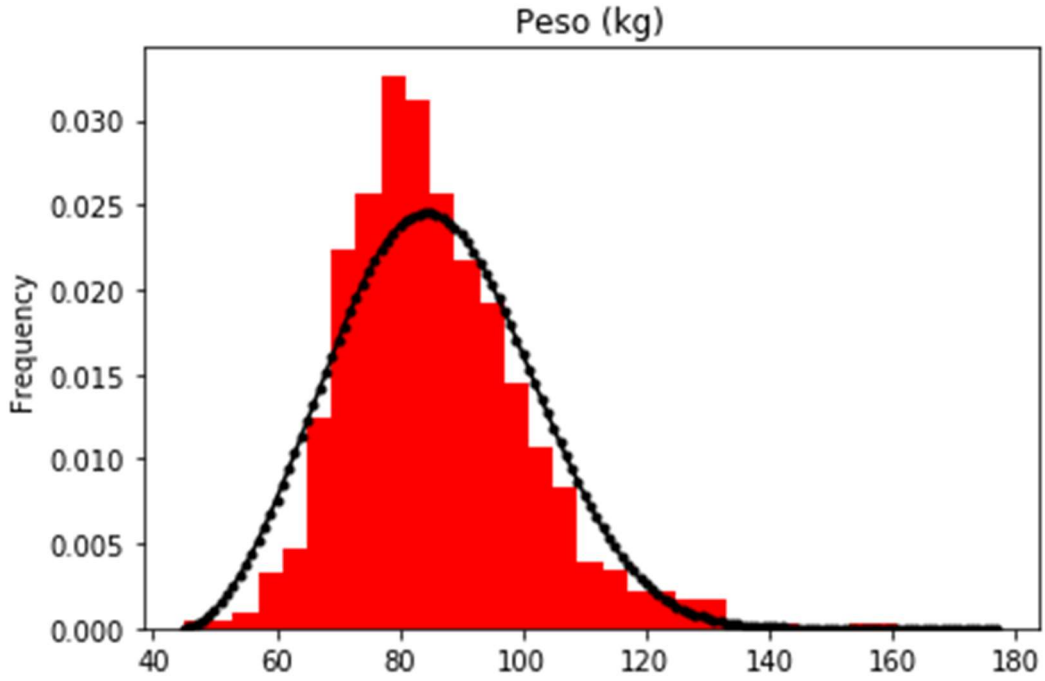


Pode-se ver com clareza como as distribuições gaussiana, lognormal e weibull se sobrepõe bem a distribuição dos dados. Somente a exponencial está desalinhada.

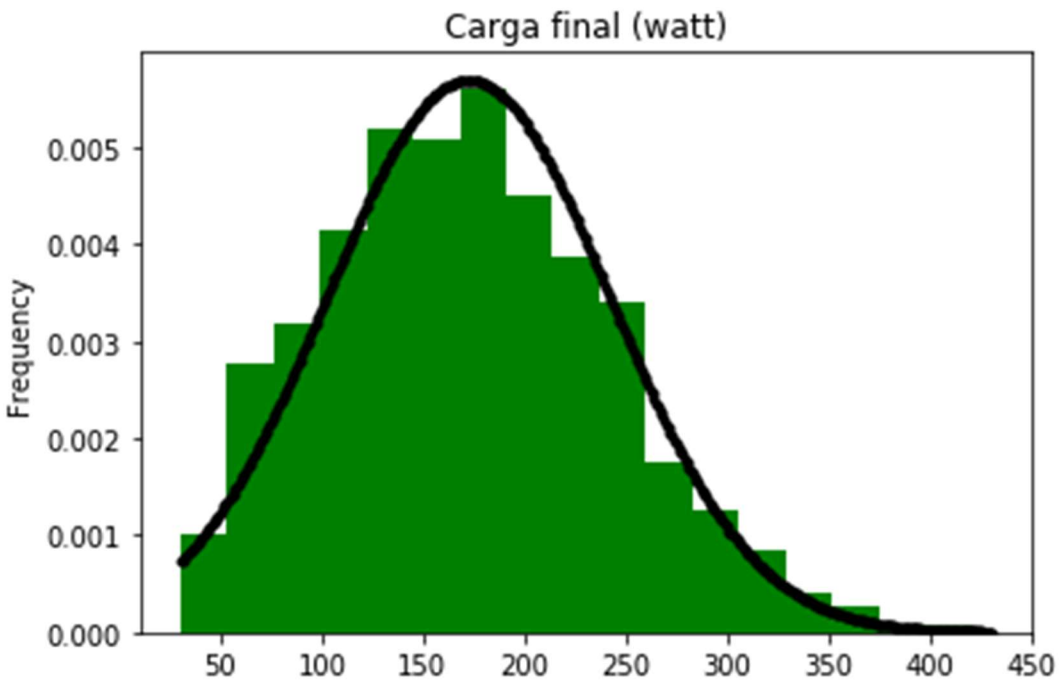

Peso (kg)
Frequency
0.030
0.025
0.020
0.015
0.010
0.005
0.000
40
60
80
100
120
140
160
180

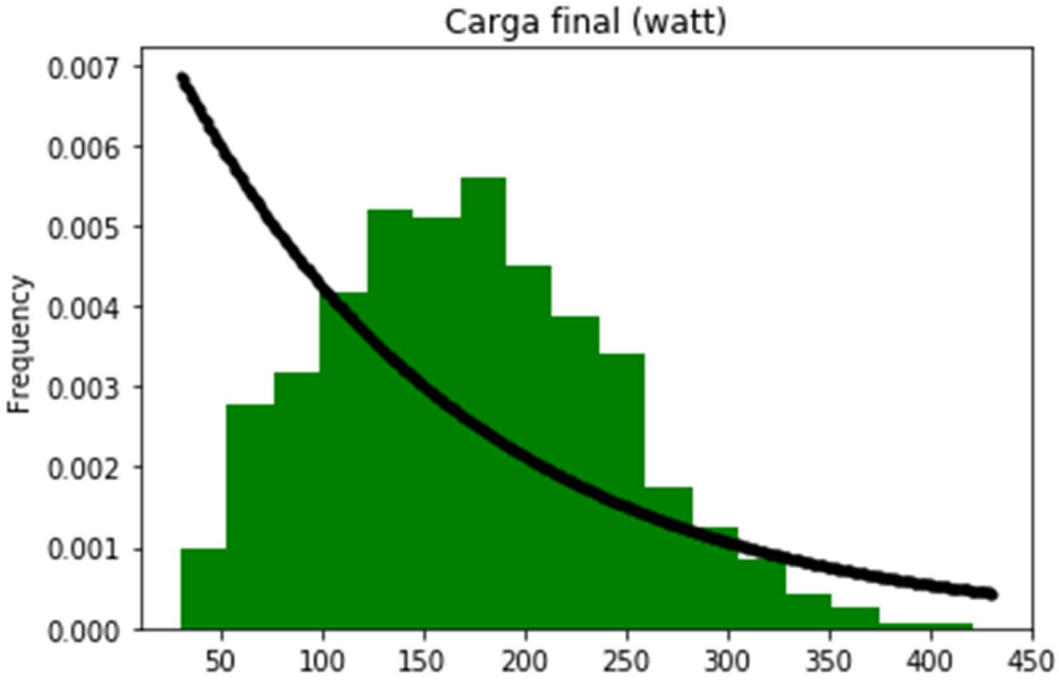

Peso (kg)
Frequency
0.030
0.025
0.020
0.015
0.010
0.005
0.000
40
60
80
100
120
140
160
180

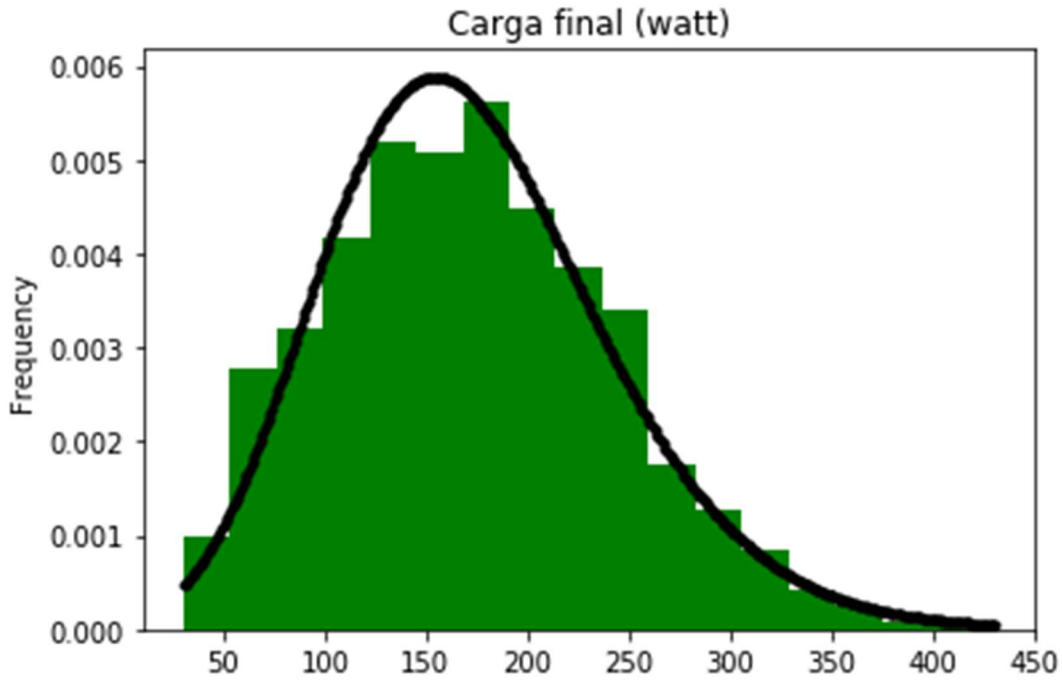


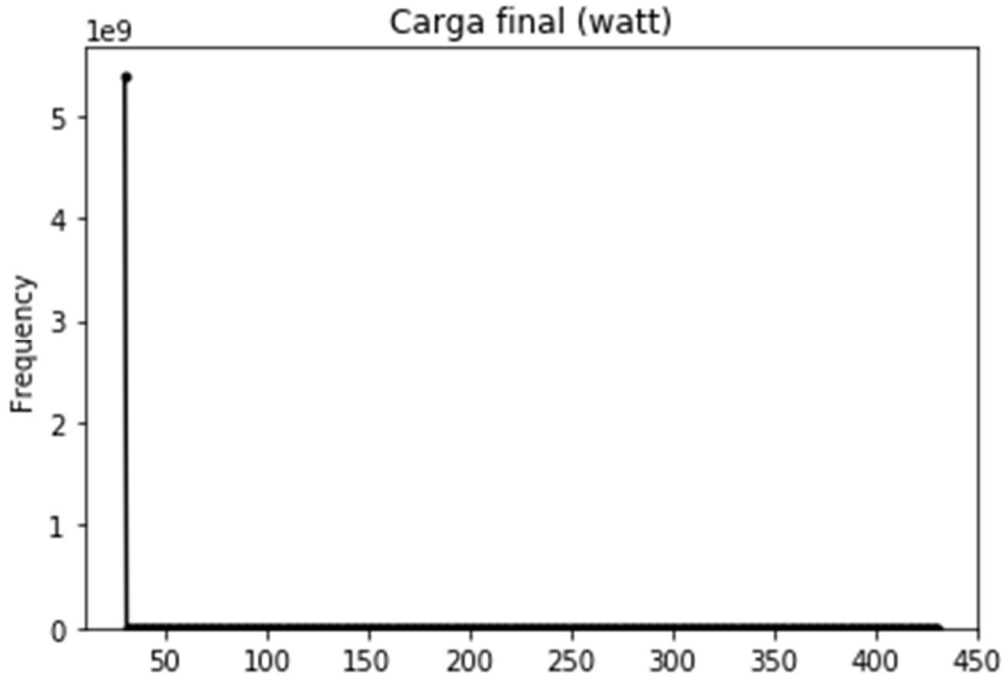


Mais uma vez pode-se ver como as distribuições gaussiana, lognormal e weibull se sobrepõe bem a distribuição dos dados, em especial a lognormal. A exponencial continua desalinhada, porém melhor na variável Peso do que na variável Idade.

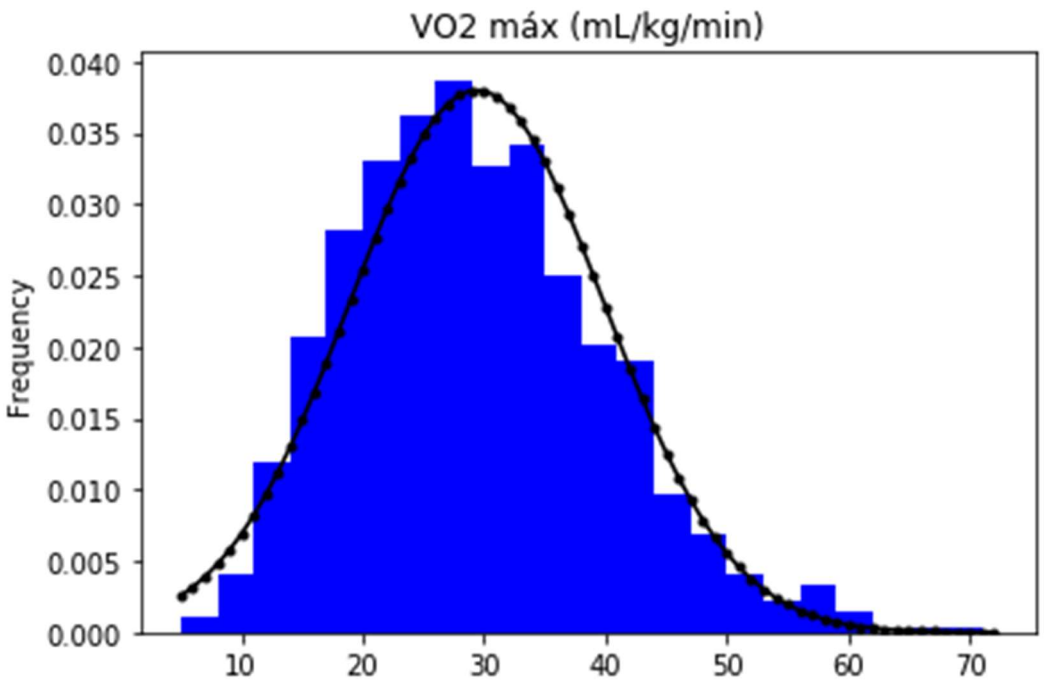

Carga final (watt)
Frequency
0.005
0.004
0.003
0.002
0.001
0.000
50
100
150
200
250
300
350
400
450

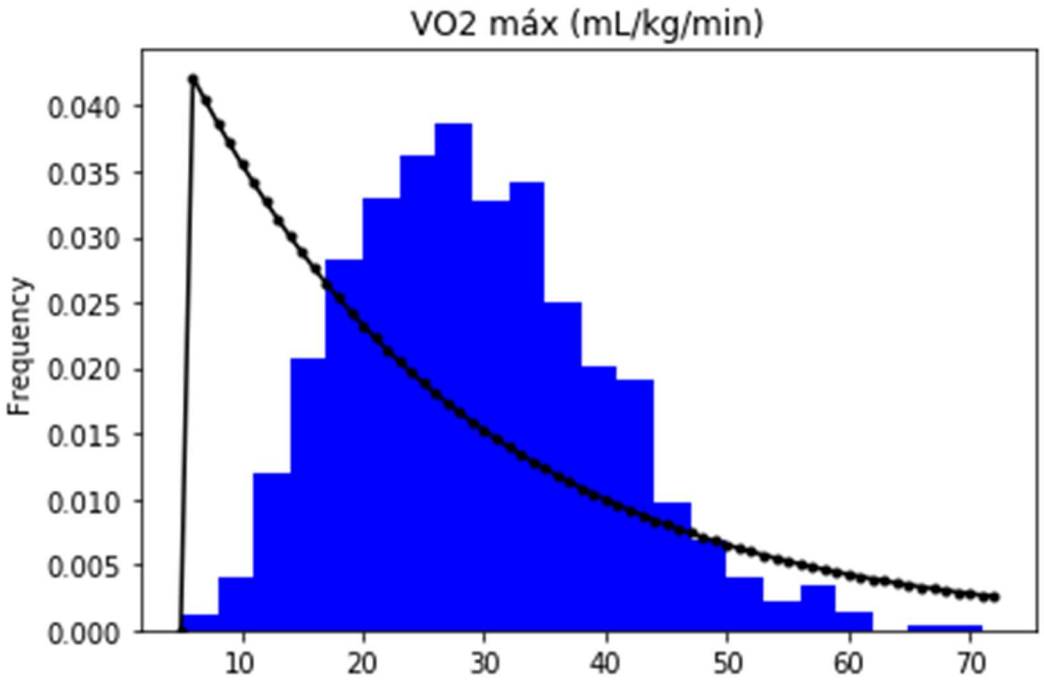

Carga final (watt)
Frequency
0.007
0.006
0.005
0.004
0.003
0.002
0.001
0.000
50
100
150
200
250
300
350
400
450

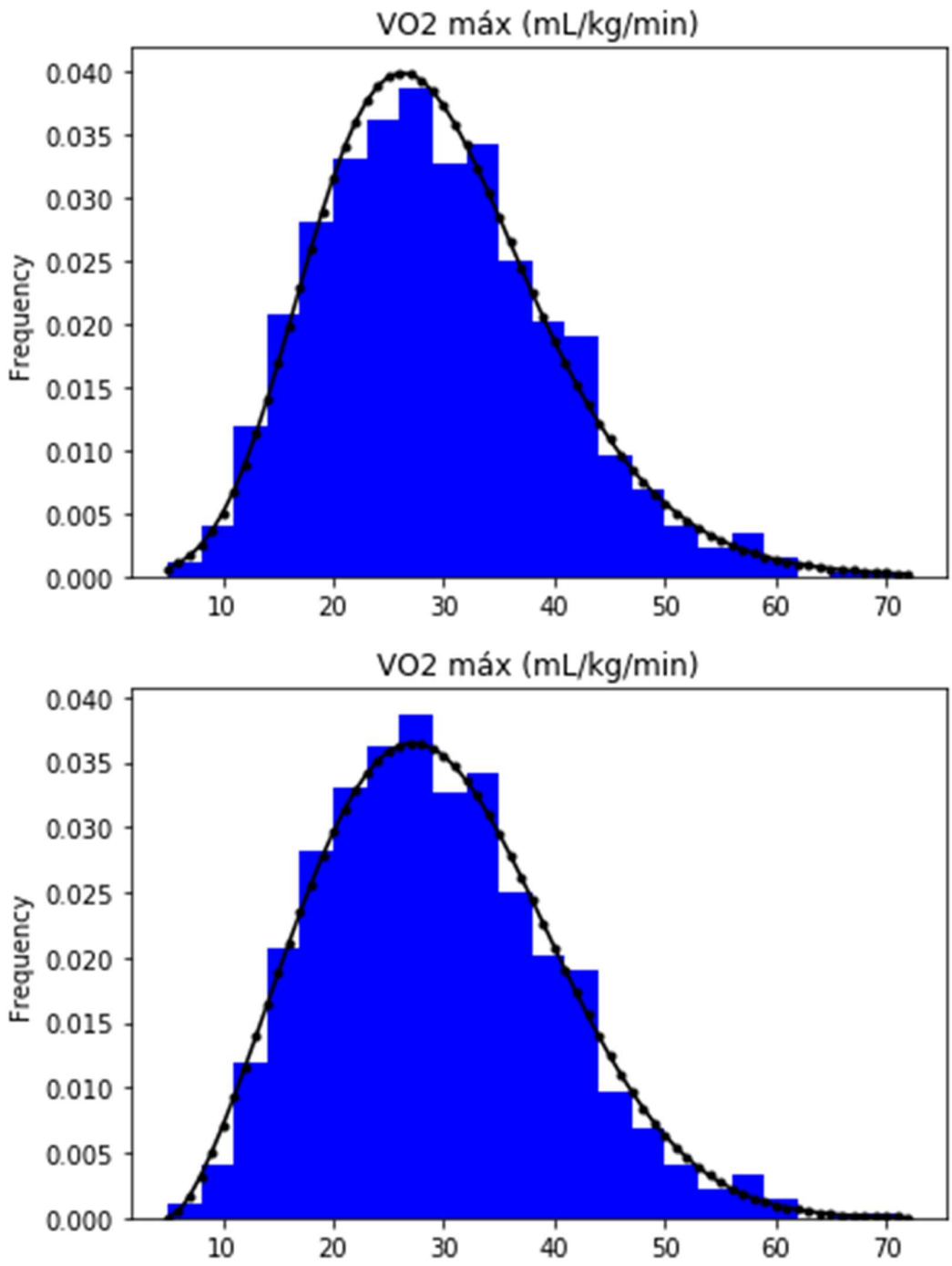




Mais uma vez pode-se ver como as distribuições gaussiana e lognormal se sobrepõe bem a distribuição dos dados. Entretanto neste último gráfico da Carga final, algum erro ocorreu na geração da sobreposição dos dados. Possivelmente um dos parâmetros pode ter sido mal calculado por falha a nível de software no tempo de execução.


VO2 máx (mL/kg/min)
Frequency
0.040
0.035
0.030
0.025
0.020
0.015
0.010
0.005
0.000
10
20
30
40
50
60
70


VO2 máx (mL/kg/min)
Frequency
0.040
0.035
0.030
0.025
0.020
0.015
0.010
0.005
0.000
10
20
30
40
50
60
70



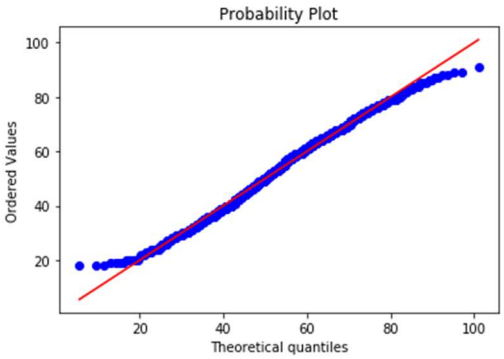
Assim como a variável Idade, para a variável VO2 máx as distribuições gaussiana, lognormal e weibull se sobrepõe bem a distribuição dos dados.

3.4. Gráfico QQplot ou ProbabilityPlot

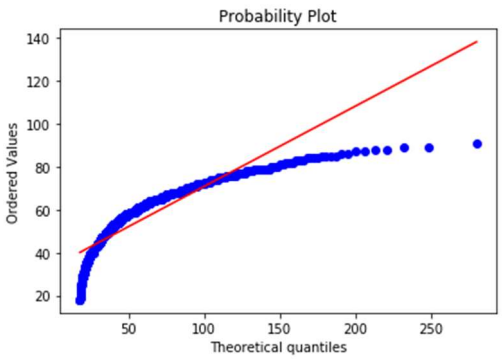
Nos gráficos de probabilidade abaixo, pode-se notar como no geral, as distribuições gaussiana, lognormal e weibull se adequam bem a modelagem dos dados.

IDADE: Vê-se com clareza como apenas a distribuição exponencial se distancia do comportamento do conjunto de dados.

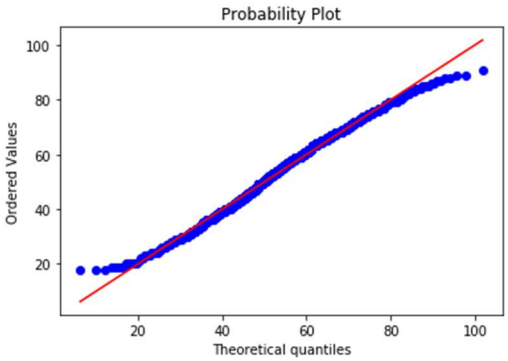
Gaussiana



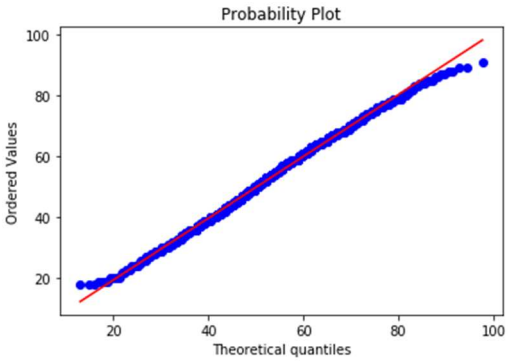
Exponencial



Lognormal

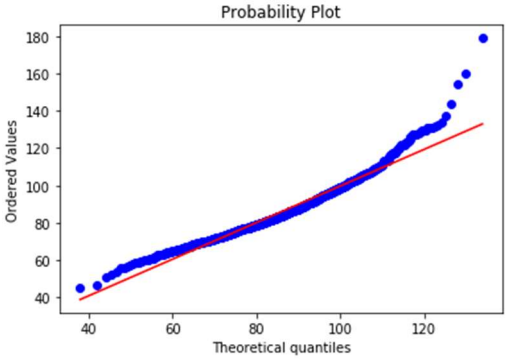


Weibull

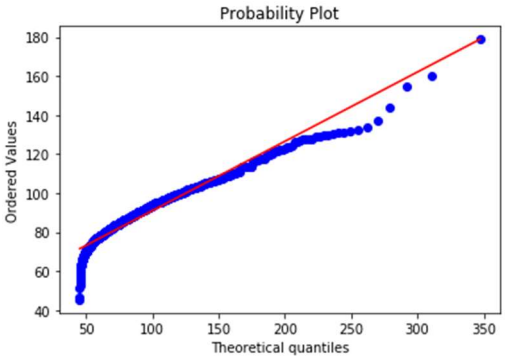


PESO : Semelhante a Idade, exceto pela distribuição Exponencial ter um rendimento maior neste caso.

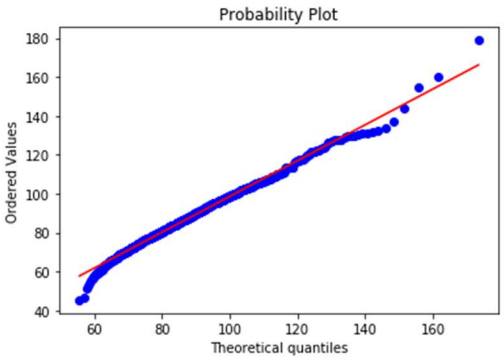
Gaussiana



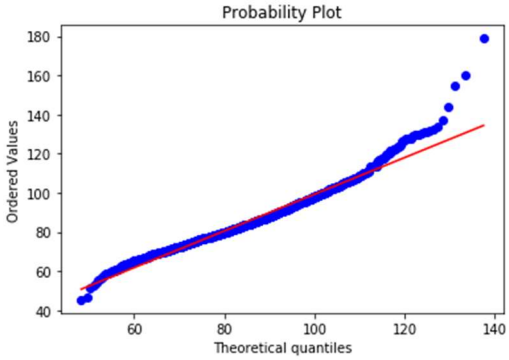
Exponencial



Lognormal

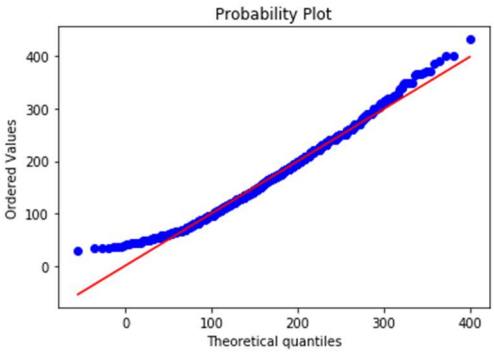


Weibull

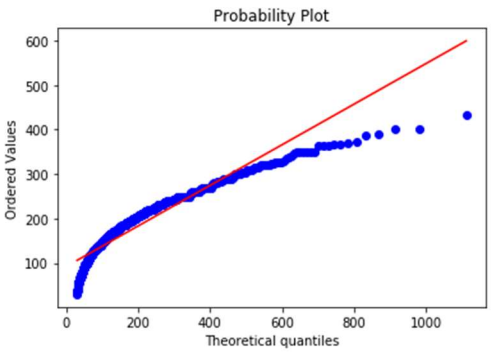


CARGA FINAL: Nota-se com clareza que a distribuição Weibull é a de menor desempenho/similaridade.

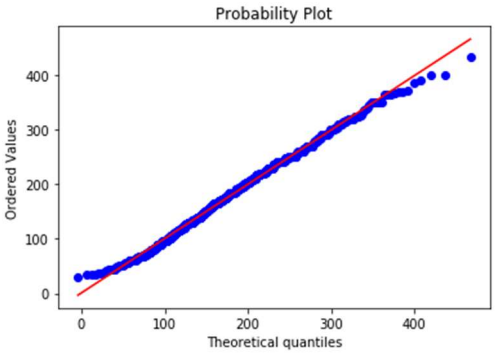
Gaussiana



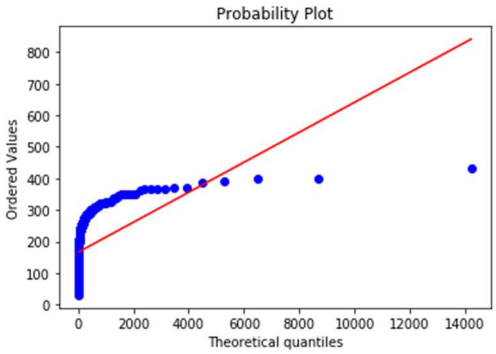
Exponencial



Lognormal

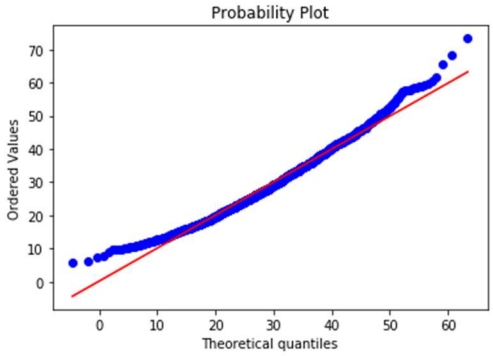


Weibull

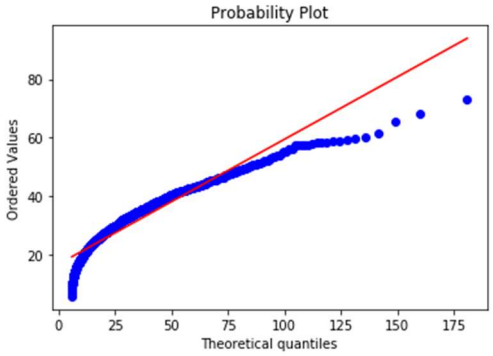


VO2 MÁX: dentre todas as variáveis, o VO2 máx aparenta possuir a melhor compatibilidade com todas as distribuições.

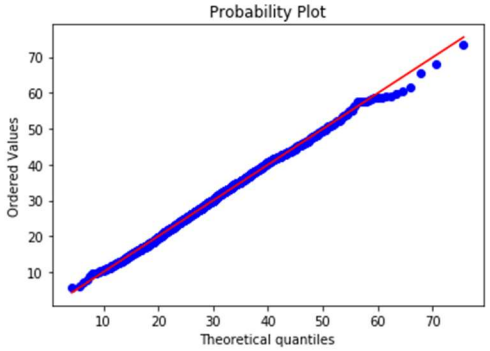
Gaussiana



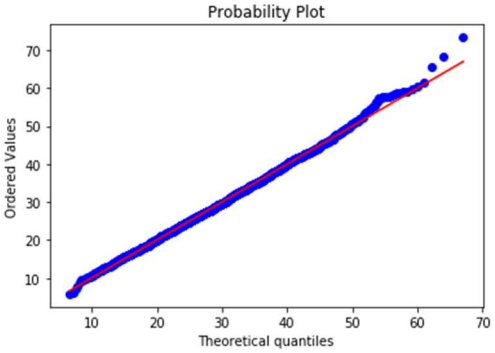
Exponencial



Lognormal



Weibull



3.5. Teste de Hipótese

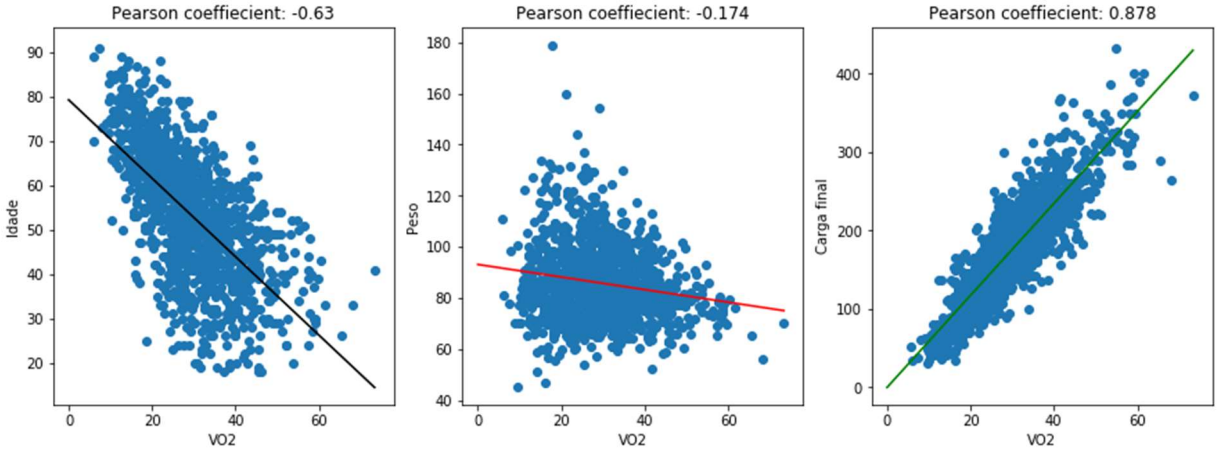
Na tabela abaixo estão dispostos os resultados dos testes de Kolmogorov-Smirnov de todas as variáveis para as distribuições estudadas neste projeto. Destacam-se a hipótese de lognormal na variável VO2 máx e Weibull na Carga final como sendo, respectivamente, as de maior semelhança e menor semelhança.

Distribuição	**Idade (stats)**	**Peso (stats)**	**Carga final (stats)**	**VO2 máximo (stats)**
Gaussiana	0.04408	0.06661	0.03923	0.04453
Exponencial	0.27666	0.35689	0.22648	0.26700
Lognormal	0.04539	0.03919	0.02888	0.01900
Weibull	0.03785	0.06066	0.75100	0.02042

3.6. Análise de dependência entre as variáveis, modelo de regressão

Pode-se notar como apenas a variável Carga final possui uma correlação significativamente alta com VO2 máx. Pode-se inferir também que a variável Idade está negativa correlacionada com VO2, e a variável Peso quase não se correlaciona com VO2.

Correlações	Correlação de Pearson a VO2 máximo
Idade	-0.63007
Peso	-0.17440
Carga final	0.878325



Deste modo, poderia ser usado um modelo de regressão linear para representar uma relação entre VO2 e Carga final.

3.7. Inferência Bayesiana

Como a variável de maior correlação com VO2 é a Carga final, a heurística utilizada neste trabalho foi de particionar os dados em quatro (4) hipóteses:

- Hipótese 1 (H 1): $0 \leq Carga\ final < 100$
- Hipótese 2 (H 2): $100 \leq Carga\ final < 200$
- Hipótese 3 (H 3): $200 \leq Carga\ final < 300$
- Hipótese 4 (H 4): $300 \leq Carga\ final < 400$

Sendo H_A a hipótese A, em que VO2 máx < 35, e sendo H_B a hipótese B, em que VO2 máx ≥ 35. Temos as seguintes tabelas de inferência.

H_A

Hipótese	Priori P(H)	Likelihood P(H_A | H)	Bayes Numerator P(H) * P(H_A | H)	Posteriori P(H | H_A)
H 1	0.16211	1.0	0.16211	0.18410
H 2	0.49744	0.90909	0.45221	0.51356
H 3	0.290955	0.36363	0.26450	0.30038
H 4	0.046928	0.03636	0.00170	0.00193

H_B

Hipótese	Priori P(H)	Likelihood P(H_B | H)	Bayes Numerator P(H) * P(H_B | H)	Posteriori P(H | H_B)
H 1	0.16211	0.0	0.0	0.0
H 2	0.49744	0.09090	0.04522	0.38686
H 3	0.29095	0.63636	0.02645	0.22627
H 4	0.04692	0.96363	0.04522	0.38686

Para realizar a previsão de um valor de VO2 que inicialmente estava contido na hipótese A, e que passaria a ser da hipótese B, $P[\, VO_2máximo \geq 35 \,|VO_2máximo < 35]$, substitui-se os valores da Priori na tabela H_B pelos valores da posteriori da tabela H_A e recalculando o numerador de Bayes e, por conseguinte, a Posteriori (a predição em si).

Tal tarefa foi realizada na tabela abaixo:

Hipótese	Priori P(H)	Likelihood P(H_A | H)	Bayes Numerator P(H) * P(H_A | H)	Posteriori P(H | H_A)
H 1	0.18410	0.0	0.0	0.0
H 2	0.51356	0.09090	0.04668	0.19476
H 3	0.30038	0.63636	0.19114	0.79751
H 4	0.00193	0.96363	0.00185	0.00767