

Versi-
color



Virginica

Setosa



Chin-
strap



Gen-
too



Adelie

アイリス VS

パンギン

furukawa

趣旨

- ・ 機械学習に使用される定番の座を狙う
PalmerPenguinsデータが2020年の夏に登場しました。
- ・ Iris vs PalmerPenguins 比較をしました。
→ どちらもすぐ精度出て練習用として使いやすい。
Penguinsのほうが特徴量の数が多いです。
- ・ プログラム言語 R を使ってみました。
→ 意外と簡単に使えました。

Iris

機械学習データの定番。

Scikit-learnにもデータセットとして含まれている。

1936年頃から使われている。

3種類の花を、4つの特徴量（がくと花びらの長さや幅）で分類する。

各50サンプルずつ、計150データ。

PalmerPenguins <https://allisonhorst.github.io/palmerpenguins/>

機械学習データの覇権を狙うデータセット。

2020年夏に公開された。

3種類のペンギンを、7つの特徴量（身体計測値、性別、計測年・場所）で分類する。

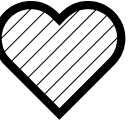
※地域や年齢などを含めたより詳しいバージョンも別である。

Adelie 152、Chinstrap 68、Gentoo 124、計344データ。

String型やNA（ごく少量）が含まれるので処理の練習にはなりそう。

Iris vs PalmerPenguins 印象

PalmerPenguins =

Iris + α + 分かりやすいかわいさ 

α = 文字列、NaN

α は別にほかのデータセットでもいいような。。

ちなみに

🐧 この絵文字：皇帝ペンギン
ピングー：皇帝ペンギン
タキシードサム：皇帝ペンギン

Suicaのペンギン：アデリーペンギン

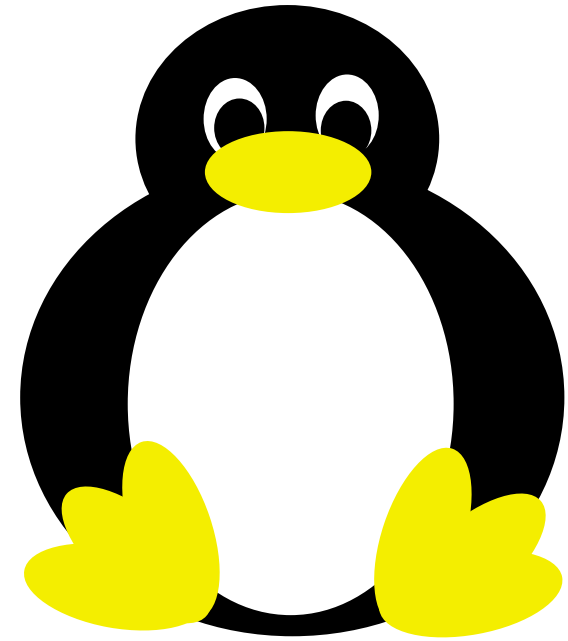
Linuxのペンギン(Tux)：アデリーペンギン？

※開発者のリーナス・トーバルズが休暇中にコガタペンギンにかまれたことが
マスコットにペンギンが選定されたきっかけらしいが、

コガタペンギンはくちばしが黒いので、マスコットの見た目とは異なる。

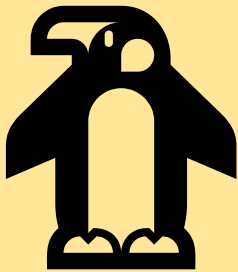
バッドばつ丸：アデリーペンギン？

バッドばつ丸のパパ：ギャングのボス（いいのだろうか）



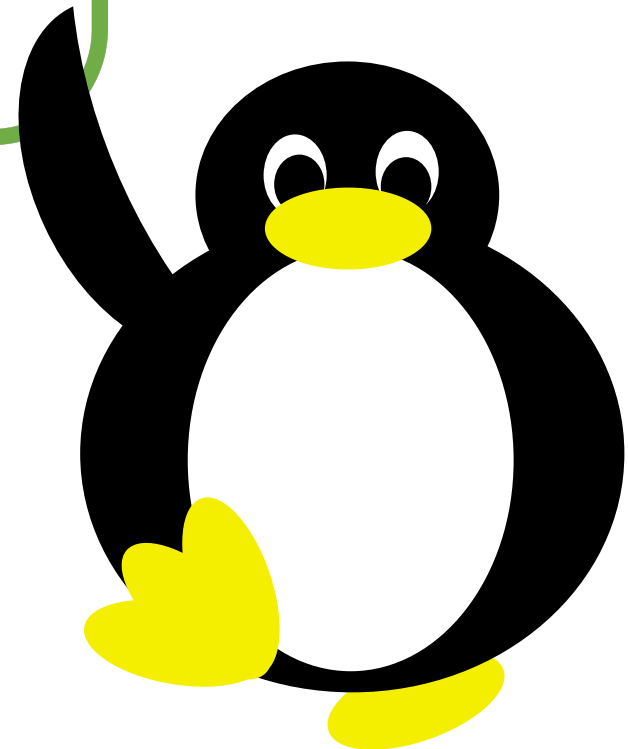
評価方法

初心者の機械学習練習用なので
簡単に一通りの練習ができるかを見る

		アイリス 	
		できる	できない
<div>ペンギン</div> 	できる	<ul style="list-style-type: none">・簡単に精度を出す・分布を見る	<ul style="list-style-type: none">・NaNの処理・文字列の処理・コード例
	できない	<ul style="list-style-type: none">・前処理なしで使う・ライブラリに入っている	—

まあね。。。。

ペンギンデータはアイリスの後発なので
アイリスでできることは
ペンギンでもできるように
なってます。



項目を比較

Iris : 5項目

sepal length (cm), sepal width (cm), petal length (cm), petal width (cm), target

PalmerPenguins : 8項目

species, island, bill_length_mm, bill_depth_mm, flipper_length_mm,

body_mass_g, sex, year

※調査の元データ（17項目）も付属している

精度：ある程度のスコアが出てほしい


それぞれ2項目だけを使って

train:test=8:2 でデータを分けてLightGBMClassifier で学習させた。

Iris testscore : 0.93

Penguin testscore : 0.92 (6/69が推論間違い)

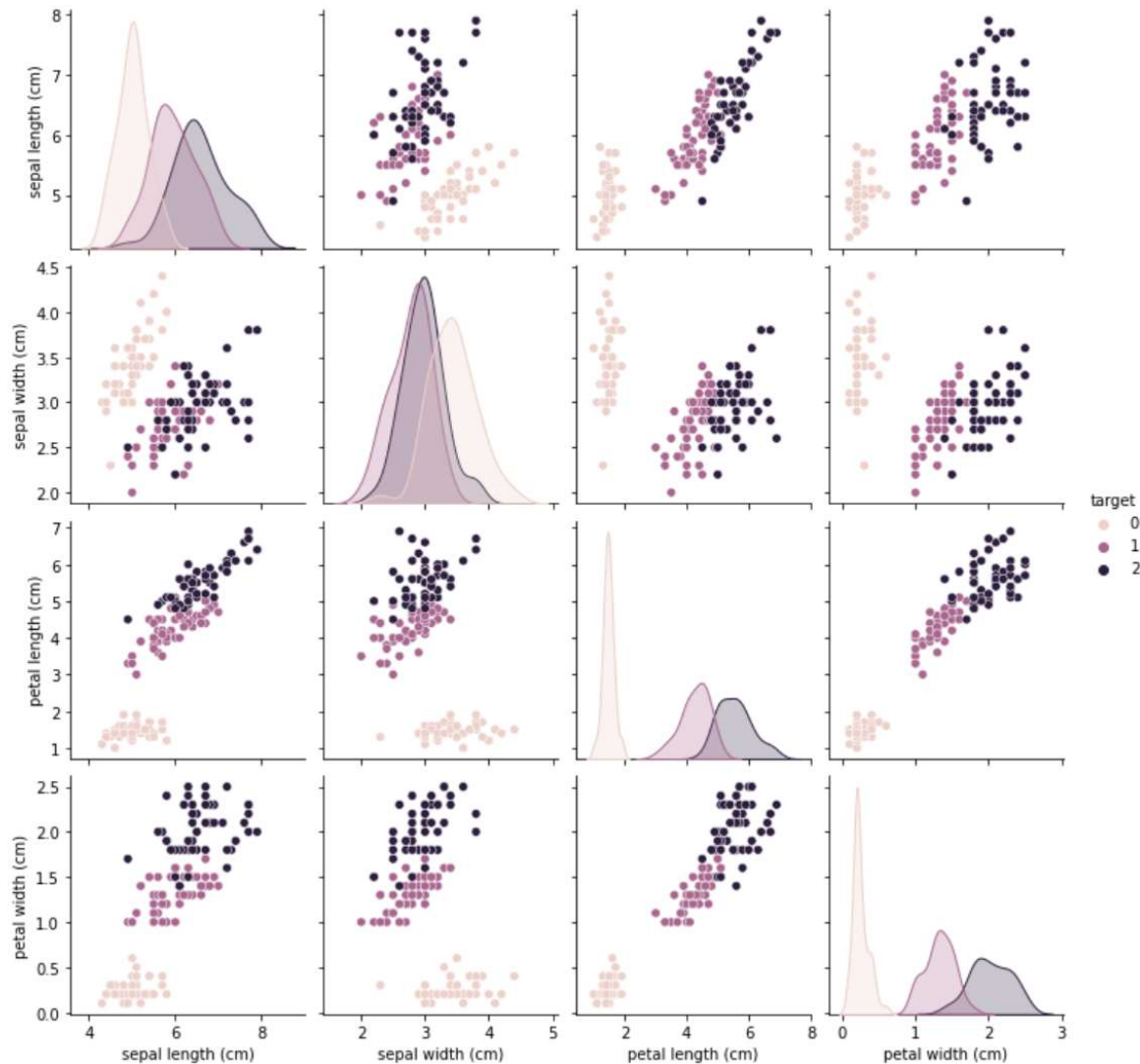
	target	target_pred	TF
13	2	0	False
14	2	1	False
15	2	1	False
35	0	1	False
38	1	0	False
54	1	0	False

 0.92なら十分。

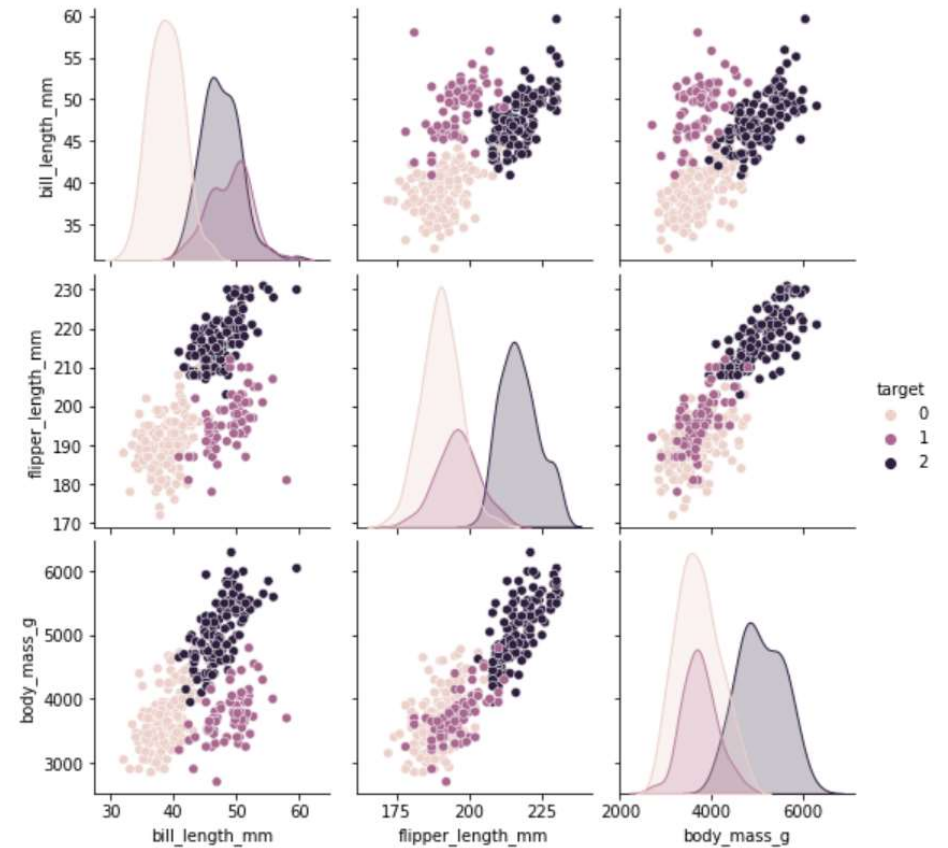
間違いには特に傾向はみられなかった

ペアプロット

Iris



PalmerPenguins



いい感じに分布している
基本的なことは一通りできそう

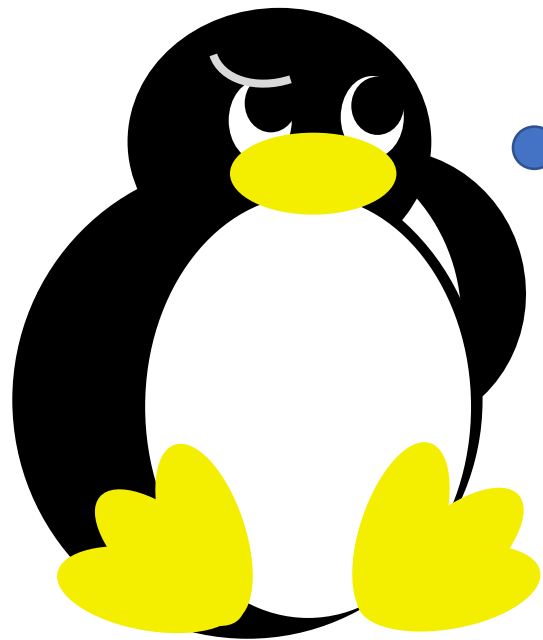
PalmerPenguinsに勝ち目はあるか？

NANの処理や文字列の処理など



PalmerPenguinsでできてIrisではできないことというのはあるが、

「それがやりたいならTitanicでいいじゃないか」という考えもありますね。

※Titanic：タイタニック号の乗客データから生存者を予測する趣旨のデータセット



Titanicデータは
悲しい気持ちになるから
“かわいい”ペンギンがいいよね
という、単なる好みの話になってしまう。。
データセットとしての利点が
いまいち説明しづらい。

	いいところ	気になるところ
ペンギン 	<ul style="list-style-type: none">・テーマがかわいい・簡単に精度が出る	<ul style="list-style-type: none">・精度上げについて工夫するところがあまりなさそう
タイタニック 	<ul style="list-style-type: none">・「客室の等級と生存率は関係するのではないか？」など、人間の推測を分類に使うことができる。←強い魅力・工夫によって精度がだいぶ変わる。	<ul style="list-style-type: none">・テーマが重い

PalmerPenguinsを盛り上げるには

フォーラム中に出た意見で、

「このペンギンデータと対になるペンギン画像セットを用意し、

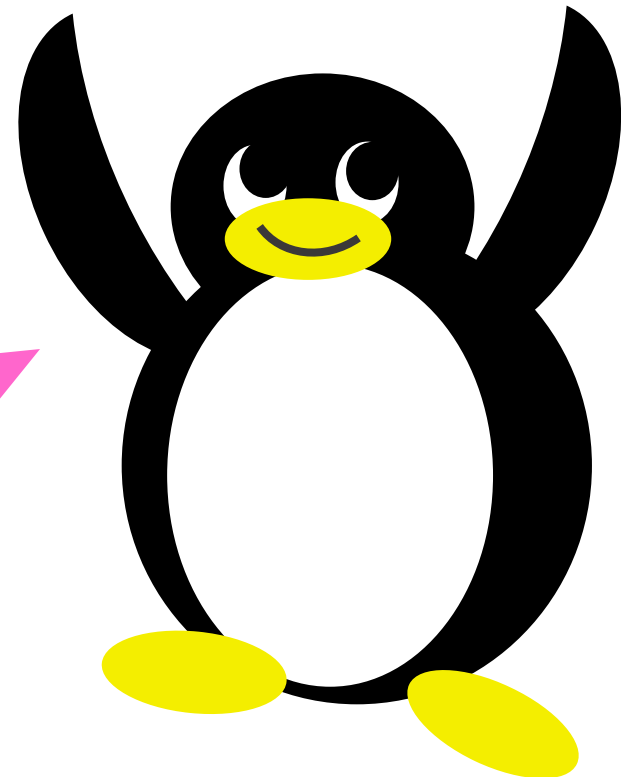
“機械学習も深層学習（画像分類・物体検出）も両方できる”

という売りにしてはどうか？」

というものがありませんでした。

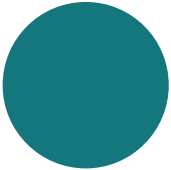
それだー！！

“機械学習”というくくりで考えると
特徴量を増やしたりする方向へいってしまうが
“初学者学習用データセット”と考えれば
機械学習も深層学習もできる、
というのはめっちゃうれしい！



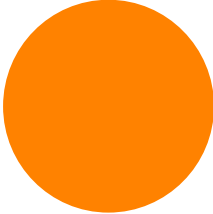
PalmerPenguinsのすてきさ

個人的に感じたPalmerPenguinsのすてきさは、その提供方法です。
そもそもテーマになっているペンギン自体がかわいいです。
コード例が複数提供されていてすぐ使えます。
そして、プロットされるときの色が指定されていて、その色がおしゃれなんです。
昔の『フラニーとズーイ』の表紙を思い出します。




公式ページのアートワークも

カラフルなペンギンのイラスト満載でかわいいです。



かわいくて、すぐ結果が出る、
そうすると、やっていることがうれしくなります。
そこが、いいなと思いました。



話変わって、気になるのが

Palmer Penguin のデータセットのチュートリアルは

R 前提です（pythonではない）。

Rでのコード例が入っています。

Rは統計に強いプログラム言語だそうです、

私は使ったことはありません。

なんとなく難しそうなイメージです。

使ってみました。

→ 意外に使いやすいです

R

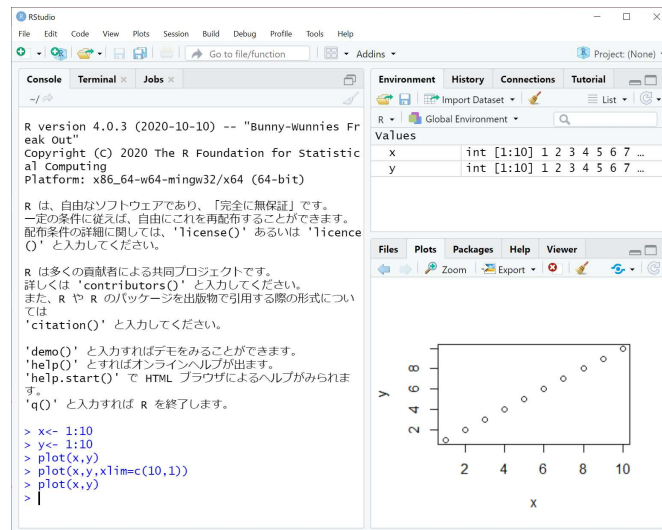
■Rとは

統計解析向けのプログラミング言語
プロットが簡単にできる

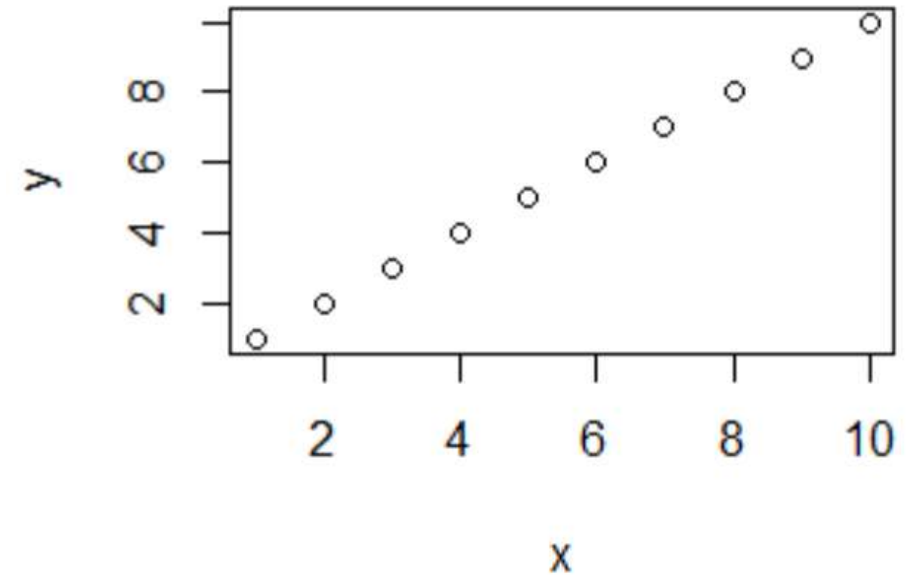
■インストール：無料でできる

Rをインストール (v4.0.3)

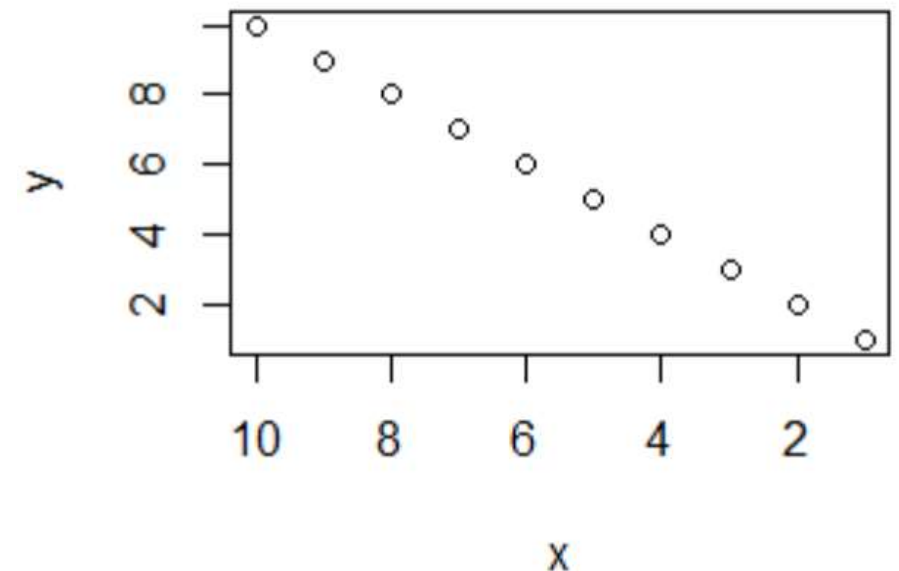
R studioをインストール



```
> x<- 1:10  
> y<- 1:10  
> plot(x,y)
```



```
> plot(x,y,xlim=c(10,1))
```



右図がコード例です。シンプル。

ColaboratoryでRを使う

上部バーより「ランタイム」→「ランタイムのタイプを変更」→R

あとは、PalmerPenguinsのチュートリアル通りに実行すれば使えます。

```
[2] install.packages("palmerpenguins")
```

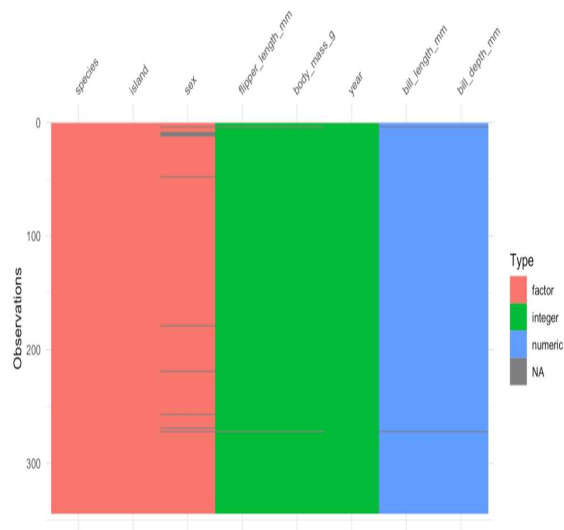
Installing package into '/usr/local/lib/R/site-library'
(as 'lib' is unspecified)

```
[3] library(palmerpenguins)  
data(package = 'palmerpenguins')
```

```
[4] head(penguins)
```

A tibble: 6 × 8

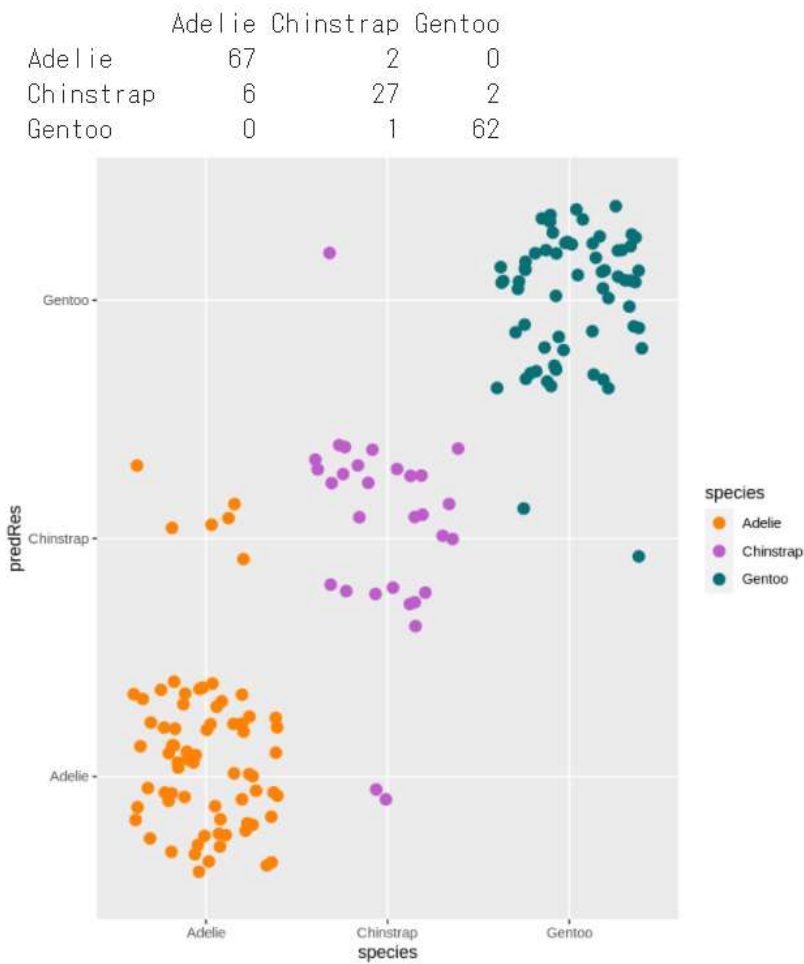
species	island	bill_length_mm	bill_depth_mm	flipper_length_mm	body_mass_g	sex	year
<fct>	<fct>	<dbl>	<dbl>	<int>	<int>	<fct>	<int>
Adelie	Torgersen	39.1	18.7	181	3750		
Adelie	Torgersen	39.5	17.4	186	3800		
Adelie	Torgersen	40.3	18.0	195	3250		



きれいです。

RでPalmerPenguins分類を推論

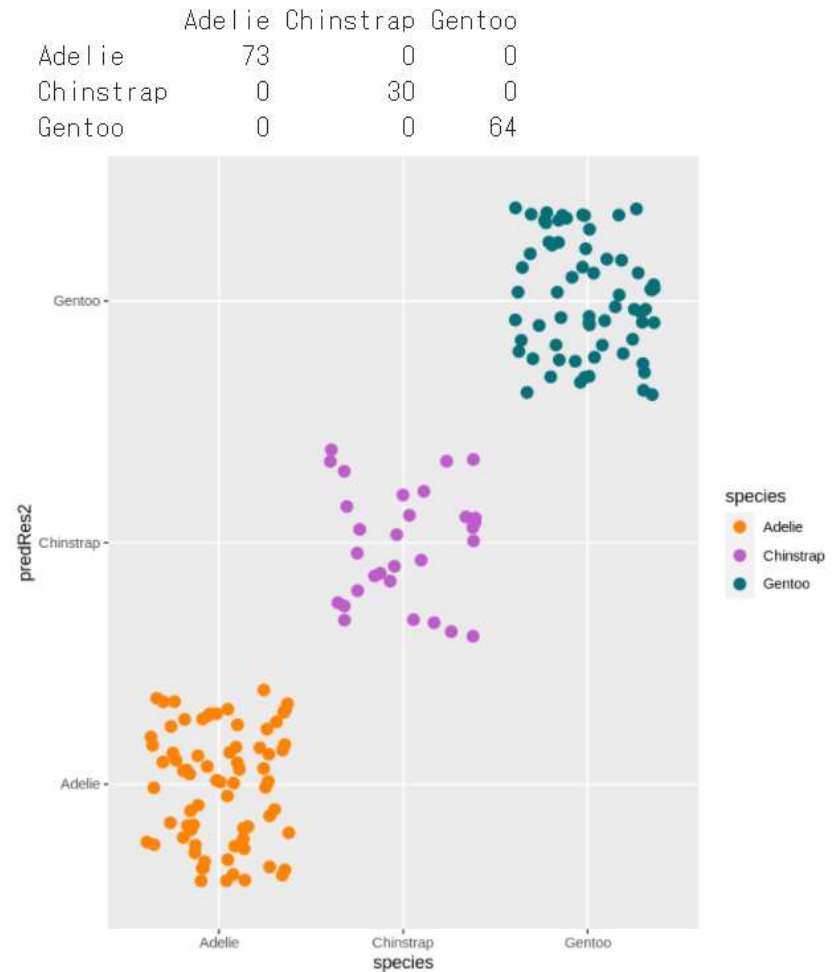
linear modeling : 0.93



$$((67+2+6+27+2+1+62)-(6+2+1+2))/((67+2+6+27+2+1+62))$$

0.934131736526946

svm : 1.0



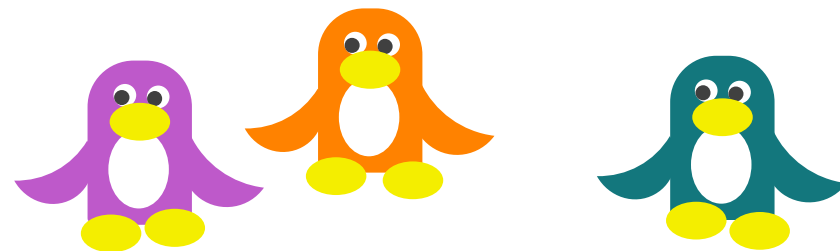
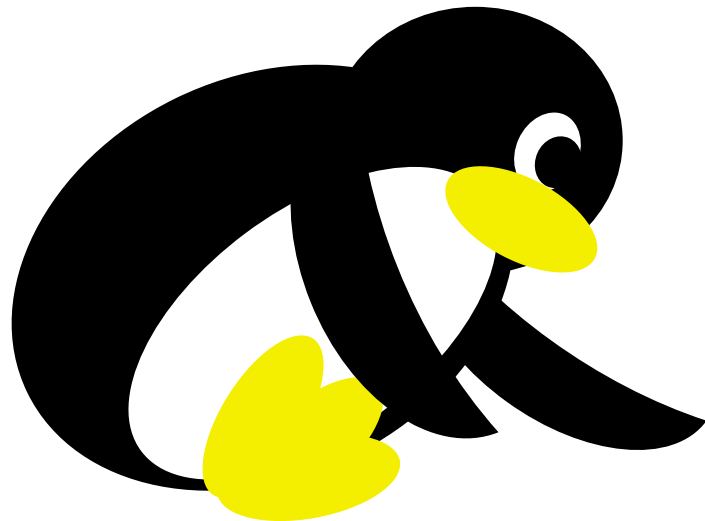
$$(73+30+64)/(73+30+64)$$

1

まとめ

PalmerPenguins自体は機械学習の勉強用としていいデータです。
サイズも小さくて扱いやすいし、テーマがかわいくて楽しいですし。

「Rを今後使いたいのか？」という質問をフォーラム時にいただきました。
現在使っているpythonは
ライブラリが豊富で機械学習以外のこともできますし、
特に不便はないです。
Rに乗り換えることはないと思います。



ぴっぴっぴ

