|  |  |  |
| --- | --- | --- |
| **Nama :**  **NIM :** ….... | c9824-logouniversitastrisakti  **Praktikum Data Analitik** | **Modul 7**  **Nama Dosen:** **Syandra Sari, S.Kom, M.Kom** |
| **Hari/Tanggal** :  .., .. November 2021 | **Nama Aslab :**   1. **Tasya Aulia (065001800022)** 2. **Shabrina Irsyadyahwati (065001800023)** |

**Praktikum 7 – REGRESI LOGISTIK**

**DESKRIPSI MODUL** : Melakukan pengujian Regresi Logistik

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| No | Elemen Kompetensi | Indikator Kinerja | Jml  Jam | hlm |
| 1 | Mampu melakukan pengujian regresi Logistik | Mampu melakukan pengujian regresi Logistik | 2 |  |

**TEORI SINGKAT**

Berbeda dengan regresi lininer yang digunakan untuk memprediksi variabel Y yang bersifat kontinu, regresi logistic digunakan untuk memprediksi variabel Y yang bersifat kategorik. Kasus regresi logistic dengan Y yang terdiri dari hanya dua kelas dinamakan *binary classification problems (****binomial logistic regression)****.* Prediktor dapat bersifat kontinu, kategorik maupun gabungan keduanya.

# ELEMEN KOMPETENSI I

|  |  |
| --- | --- |
| Deskripsi : | Dapat melakukan pengujian regresi Logistik |
| Kompetensi Dasar : | Mampu melakukan pengujian regresi Logistik |

**Kasus 1 : Titanic Dataset**

Gunakan data berikut ini untuk membangun model prediktif. Berikan interpretasi atas setiap output yang dihasilkan. Mulailah analisis dengan membuat tabulasi silang setiap predictor yang bersifat kategorik dengan respon (Y).

### Data Dictionary

|  |  |  |
| --- | --- | --- |
| **Variable** | **Definition** | **Key** |
| survival | Survival | 0 = No, 1 = Yes |
| pclass | Ticket class | 1 = 1st, 2 = 2nd, 3 = 3rd |
| sex | Sex |  |
| Age | Age in years |  |
| sibsp | # of siblings / spouses aboard the Titanic |  |
| parch | # of parents / children aboard the Titanic |  |
| ticket | Ticket number |  |
| fare | Passenger fare |  |
| cabin | Cabin number |  |
| embarked | Port of Embarkation | C = Cherbourg, Q = Queenstown, S = Southampton |

**REGRESI LOGISTIK**

|  |
| --- |
| > databaru=read.delim("clipboard")  > str(databaru)  > sampel1<-sample(1:nrow(databaru),0.75\*nrow(databaru))  > traininglogistik<-data.frame(databaru)[sampel1,]  > testinglogistik<-data.frame(databaru)[-sampel1,]  > modellogistik=glm(Survived~.,data=traininglogistik,family = binomial)  > summary(modellogistik) |

**#MELAKUKAN PREDIKSI**

|  |
| --- |
| > prediksilogistik=predict(modellogistik,testinglogistik)  > pred\_logreg<-as.numeric(prediksilogistik>.5)  > tabel\_logreg<-table(pred\_logreg,testinglogistik$Survived)  > tabel\_logreg    > confusionMatrix(pred\_logreg,testinglogistik$Survived,positive = "1") |

Output :

> datashab=read.csv("C:/Users/ACER/Downloads/Titanic.csv")

> str(datashab)

'data.frame': 680 obs. of 8 variables:

$ PassengerId: int 1 2 3 4 5 7 8 9 10 11 ...

$ Survived : int 0 1 1 1 0 0 0 1 1 1 ...

$ Pclass : int 3 1 3 1 3 1 3 3 2 3 ...

$ Sex : Factor w/ 2 levels "female","male": 2 1 1 1 2 2 2 1 1 1 ...

$ Age : num 22 38 26 35 35 54 2 27 14 4 ...

$ SibSp : int 1 1 0 1 0 0 3 0 1 1 ...

$ Parch : int 0 0 0 0 0 0 1 2 0 1 ...

$ Fare : num 7.25 71.28 7.92 53.1 8.05 ...

> sampel1<-sample(1:nrow(datashab),0.75\*nrow(datashab))

> traininglogistik<-data.frame(datashab)[sampel1,]

> testinglogistik<-data.frame(datashab)[-sampel1,]

> modellogistik=glm(Survived~.,data=traininglogistik,family = binomial)

> summary(modellogistik)

Call:

glm(formula = Survived ~ ., family = binomial, data = traininglogistik)

Deviance Residuals:

Min 1Q Median 3Q Max

-2.7117 -0.6532 -0.3866 0.6375 2.5366

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) 4.7615506 0.7310120 6.514 7.34e-11 \*\*\*

PassengerId 0.0010699 0.0004786 2.236 0.0254 \*

Pclass -1.2331735 0.1996293 -6.177 6.52e-10 \*\*\*

Sexmale -2.6332408 0.2648726 -9.942 < 2e-16 \*\*\*

Age -0.0400990 0.0095800 -4.186 2.84e-05 \*\*\*

SibSp -0.3844041 0.1552772 -2.476 0.0133 \*

Parch -0.1785689 0.1641924 -1.088 0.2768

Fare 0.0041282 0.0035013 1.179 0.2384

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 683.97 on 509 degrees of freedom

Residual deviance: 452.14 on 502 degrees of freedom

AIC: 468.14

Number of Fisher Scoring iterations: 5

> prediksilogistik=predict(modellogistik,testinglogistik)

> pred\_logreg<-as.numeric(prediksilogistik>.5)

> tabel\_logreg<-table(pred\_logreg,testinglogistik$Survived)

> tabel\_logreg

pred\_logreg 0 1

0 90 32

1. 7 41

> confusionMatrix(table(pred\_logreg,testinglogistik$Survived),positive = "1")

Confusion Matrix and Statistics

pred\_logreg 0 1

0 90 27

1 9 44

Accuracy : 0.7882

95% CI : (0.7191, 0.8471)

No Information Rate : 0.5824

P-Value [Acc > NIR] : 1.18e-08

Kappa : 0.5485

Mcnemar's Test P-Value : 0.004607

Sensitivity : 0.6197

Specificity : 0.9091

Pos Pred Value : 0.8302

Neg Pred Value : 0.7692

Prevalence : 0.4176

Detection Rate : 0.2588

Detection Prevalence : 0.3118

Balanced Accuracy : 0.7644

'Positive' Class : 1

**ELEMEN KOMPETENSI II**

|  |  |
| --- | --- |
| Deskripsi : | Dapat melakukan pengujian regresi Logistik dengan dataset Iris |
| Kompetensi Dasar : | Mampu melakukan pengujian regresi Logistik dengan Dataset Iris |

# Kasus 2. The iris data set (species virginica and versicolor only)

Gunakan data berikut ini untuk membangun model prediktif. Berikan interpretasi atas setiap output yang dihasilkan. Mulailah analisis dengan membuat tabulasi silang setiap predictor yang bersifat kategorik dengan respon (Y).

# make a reduced iris data set that only contains virginica and versicolor species

> library(dplyr)

> iris.small <- filter(iris, Species %in% c("virginica", "versicolor"))

# logistic regression

> glm.out <- glm(Species ~ Sepal.Width + Sepal.Length + Petal.Width + Petal.Length,

+ data = iris.small,

+ family = binomial) # family = binomial required for logistic regression

> summary(glm.out)

> exp(coef(glm.out))

> glm.out <- glm(Species ~ Sepal.Width + Petal.Width + Petal.Length,

+ data = iris.small,

+ family = binomial)

> exp(coef(glm.out))

Output :

> #tugas2

> # make a reduced iris data set that only contains virginica and versicolor species

> library(dplyr)

> iris.small <- filter(iris, Species %in% c("virginica", "versicolor"))

> # logistic regression

> glm.out <- glm(Species ~ Sepal.Width + Sepal.Length + Petal.Width + Petal.Length, data = iris.small, family = binomial) # family = binomial required for logistic regression

> summary(glm.out)

Call:

glm(formula = Species ~ Sepal.Width + Sepal.Length + Petal.Width +

Petal.Length, family = binomial, data = iris.small)

Deviance Residuals:

Min 1Q Median 3Q Max

-2.01105 -0.00541 -0.00001 0.00677 1.78065

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) -42.638 25.707 -1.659 0.0972 .

Sepal.Width -6.681 4.480 -1.491 0.1359

Sepal.Length -2.465 2.394 -1.030 0.3032

Petal.Width 18.286 9.743 1.877 0.0605 .

Petal.Length 9.429 4.737 1.991 0.0465 \*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 138.629 on 99 degrees of freedom

Residual deviance: 11.899 on 95 degrees of freedom

AIC: 21.899

Number of Fisher Scoring iterations: 10

> exp(coef(glm.out))

(Intercept) Sepal.Width Sepal.Length Petal.Width Petal.Length

3.038345e-19 1.254665e-03 8.499013e-02 8.741145e+07 1.244887e+04

> glm.out <- glm(Species ~ Sepal.Width + Petal.Width + Petal.Length, data = iris.small, family = binomial)

Warning message:

glm.fit: fitted probabilities numerically 0 or 1 occurred

> exp(coef(glm.out))

(Intercept) Sepal.Width Petal.Width Petal.Length

1.138872e-22 2.303132e-04 2.026532e+09 2.629495e+03

**TUGAS**

Make a plot of the fitted probability as a function of the linear predictor, colored by species identity. Hint: you will have to make a new data frame combining data from the fitted model with data from the *iris.small* data frame.

> lr\_data <- data.frame(predictor=glm.out$linear.predictors, prob=glm.out$fitted.values, Species=iris.small$Species)

> ggplot(lr\_data, aes(x=predictor, y=prob, color=Species)) + geom\_point()

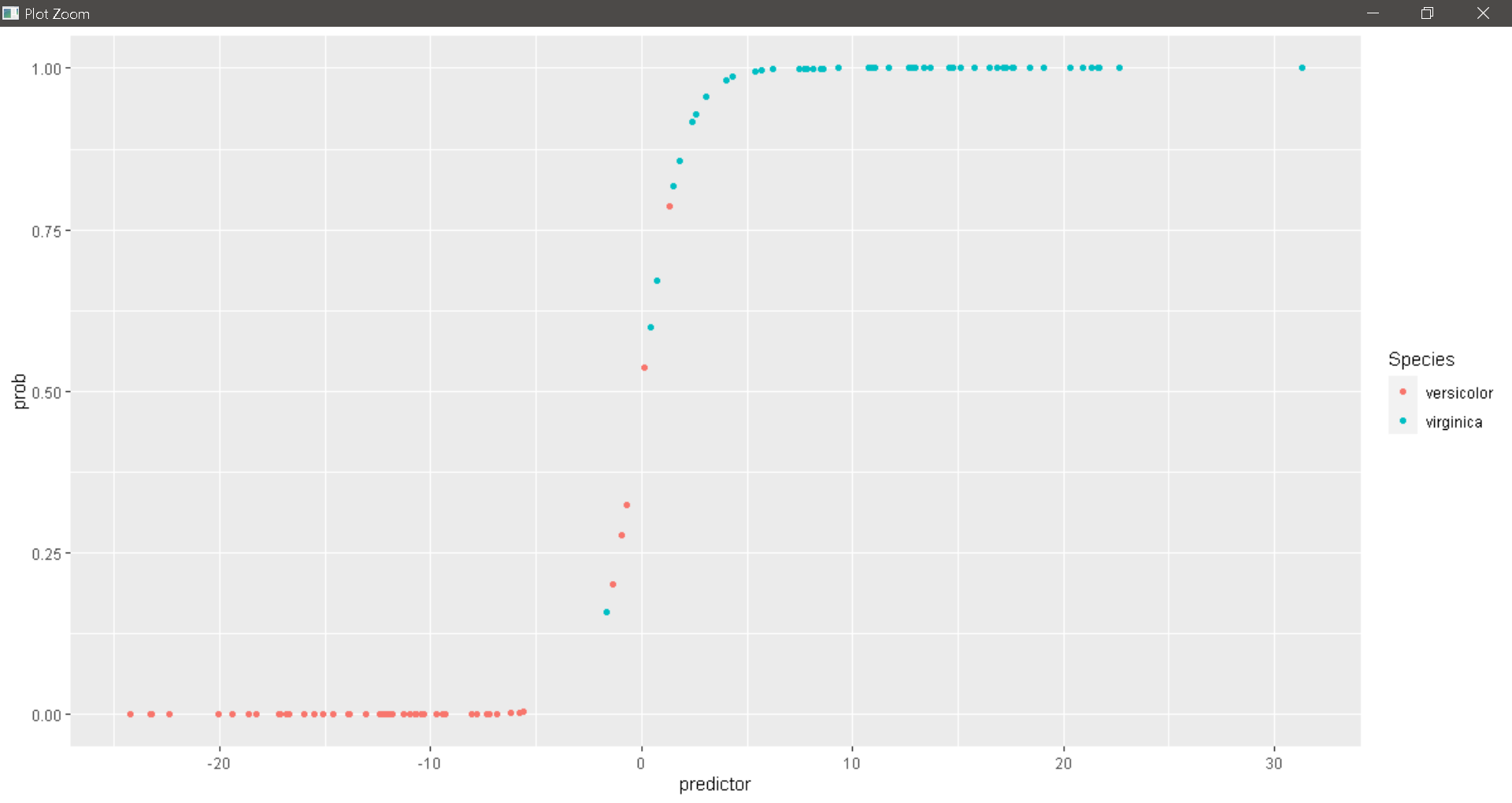
> library(ggplot2)

Warning message:

package ‘ggplot2’ was built under R version 3.6.3

> lr\_data <- data.frame(predictor=glm.out$linear.predictors, prob=glm.out$fitted.values, Species=iris.small$Species)

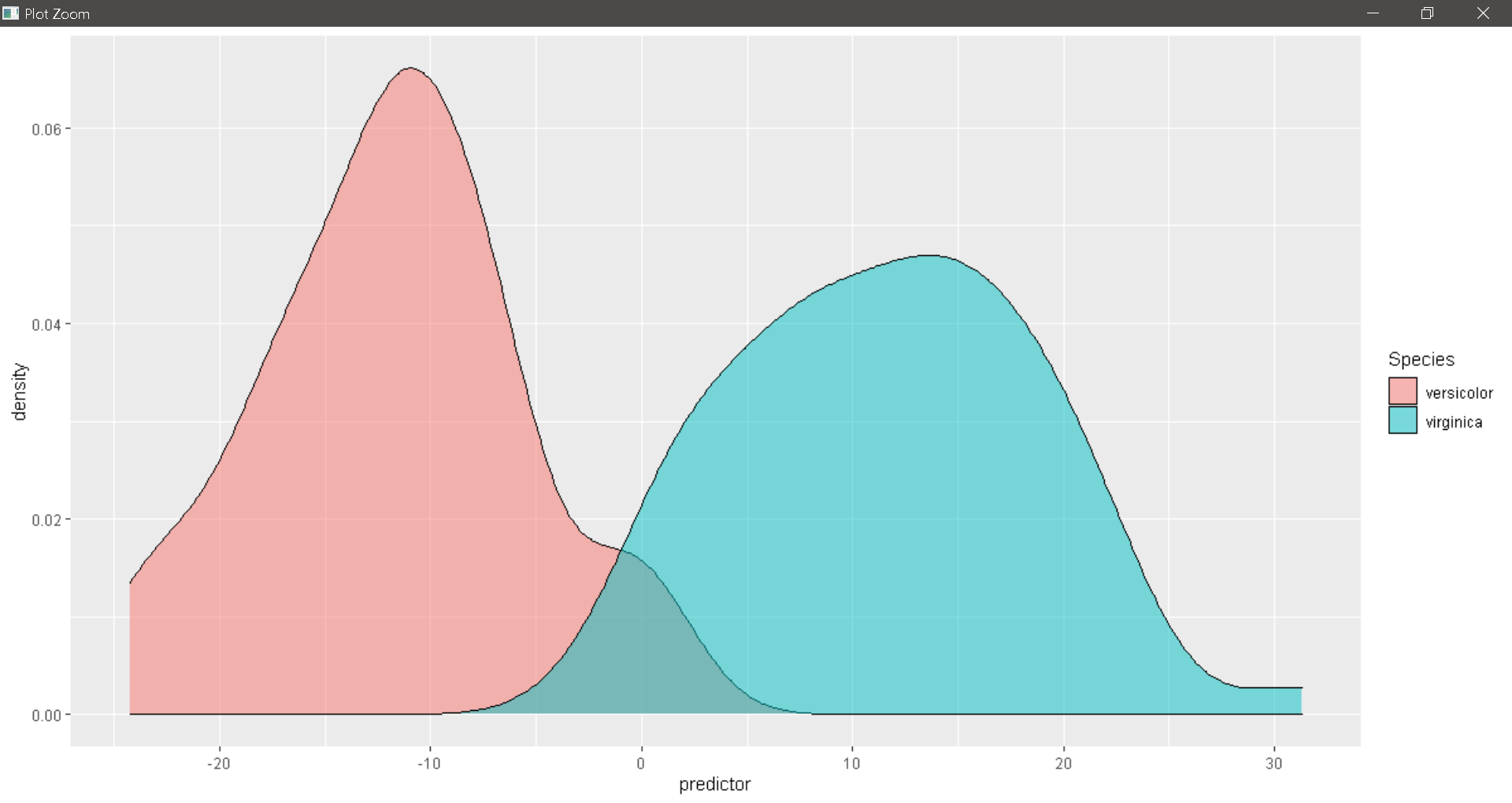
> ggplot(lr\_data, aes(x=predictor, y=prob, color=Species)) + geom\_point()



Make a density plot that shows how the two species are separated by the linear predictor.

ggplot(lr\_data, aes(x=predictor, fill=Species)) + geom\_density(alpha=.5)

> ggplot(lr\_data, aes(x=predictor, fill=Species)) + geom\_density(alpha=.5)



Assume you have obtained samples from three plants, with measurements as listed below. Predict the likelihood that each of these plants belongs to the species virginica.

> plant1 <- data.frame(Sepal.Length=6.4, Sepal.Width=2.8, Petal.Length=4.6, Petal.Width=1.8)

> plant2 <- data.frame(Sepal.Length=6.3, Sepal.Width=2.5, Petal.Length=4.1, Petal.Width=1.7)

> plant3 <- data.frame(Sepal.Length=6.7, Sepal.Width=3.3, Petal.Length=5.2, Petal.Width=2.3)

> predict(glm.out, plant1, type="response")

> predict(glm.out, plant2, type="response")

> predict(glm.out, plant3, type="response")

> plant1 <- data.frame(Sepal.Length=6.4, Sepal.Width=2.8, Petal.Length=4.6, Petal.Width=1.8)

> plant2 <- data.frame(Sepal.Length=6.3, Sepal.Width=2.5, Petal.Length=4.1, Petal.Width=1.7)

> plant3 <- data.frame(Sepal.Length=6.7, Sepal.Width=3.3, Petal.Length=5.2, Petal.Width=2.3)

> predict(glm.out, plant1, type="response")

1

0.6934611

> predict(glm.out, plant2, type="response")

1

0.06002675

> predict(glm.out, plant3, type="response")

1

0.9999943

Pick a cutoff predictor value at which you would decide that a specimen belongs to virginica rather than versicolor. Calculate how many virginicas you call correctly and how many incorrectly given that choice.

> cutoff <- 0

> virg\_true <- sum(lr\_data$predictor > cutoff & lr\_data$Species=="virginica")

> virg\_false <- sum(lr\_data$predictor <= cutoff & lr\_data$Species=="virginica")

> virg\_true

> virg\_false

> cutoff <- 0

> virg\_true <- sum(lr\_data$predictor > cutoff & lr\_data$Species=="virginica")

> virg\_false <- sum(lr\_data$predictor <= cutoff & lr\_data$Species=="virginica")

> virg\_true

[1] 49

> virg\_false

[1] 1

Now do the same calculation for versicolor.

> vers\_true <- sum(lr\_data$predictor <= cutoff & lr\_data$Species=="versicolor")

> vers\_false <- sum(lr\_data$predictor > cutoff & lr\_data$Species=="versicolor")

> vers\_true

> vers\_false

|  |
| --- |
| > vers\_true <- sum(lr\_data$predictor <= cutoff & lr\_data$Species=="versicolor")  > vers\_false <- sum(lr\_data$predictor > cutoff & lr\_data$Species=="versicolor")  > vers\_true  [1] 48  > vers\_false  [1] 2 |
|  |
|  |

If we define a call of virginica as a positive and a call of versicolor as a negative, what are the true positive rate (sensitivity, true positives divided by all possible positives) and the true negative rate (specificity, true negatives divided by all possible negatives) in your analysis?

> tp <- virg\_true/(virg\_true + virg\_false)

> tn <- vers\_true/(vers\_true + vers\_false)

> tp

> tn

> tp <- virg\_true/(virg\_true + virg\_false)

> tn <- vers\_true/(vers\_true + vers\_false)

> tp

[1] 0.98

> tn

[1] 0.96

Sumber : [**http://wilkelab.org/classes/SDS348/2015\_spring\_worksheets/class11\_solutions.html**](http://wilkelab.org/classes/SDS348/2015_spring_worksheets/class11_solutions.html)

**1.** **Cek List**

|  |  |  |  |
| --- | --- | --- | --- |
| **No** | **Elemen Kompetensi** | **Penyelesaian** | |
| **Selesai** | **Tidak** |
| **1** | **Elemen Kompetensi I**  Dapat melakukan pengujian regresi Logistik. |  |  |
| **2** | **Elemen Kompetensi II**  Dapat melakukan pengujian regresi Logistik dengan dataset Iris. |  |  |

**2.** **Form Umpan Balik**

|  |  |  |
| --- | --- | --- |
| **Elemen Kompetensi** | **Waktu Pengerjaan** | **Kriteria** |
| **Elemen Kompetensi I**  Dapat melakukan pengujian regresi Logistik. |  |  |
| **Elemen Kompetensi II**  Dapat melakukan pengujian regresi Logistik dengan dataset Iris |  |  |

Kriteria

1.Sangat Menarik

2.Cukup Menarik

3.Kurang Menarik

4.Sangat Kurang Menarik