

|  |
| --- |
| **FACULTAD DE MATEMÁTICAS** |
| **Minería de Datos** |

**Unidad III:Métodos de Clasificación**

**ADA 7: *Clasificación Bayesiana.***

*Licenciatura en Actuaria.*

**Integrantes:**

- Álvarez Herrera Samantha

- Ciau Puga Abigail

- Colonia Espinosa Cindy

- Fernández Caro Frida

- Padilla Jiménez Meybor

- Sobrino Bermejo Samantha

**M.C. Ernesto Guerrero Lara**

Los datos se encuentran en la base Titanic.

Información de la base de datos:

|  |  |
| --- | --- |
| Número de observaciones | 2201 |
| Número de Variables | 4 |

La base tiene 2201 observaciones y en ella se registran los siguientes atributos: Class, Sex y Age. La variable de clasificación es “**Sobrevivir**”, que indica si la persona sobrevivió o no.

Para el tamaño de la muestra de entrenamiento se toma la proporción 80-20, de manera que, para cada ejercicio, se utilizan 1760 datos de entrenamiento, por lo que nuestro grupo de prueba será de 441 datos.

**Modelo de dos variables**

**Clase-Sexo**

Se ajusta el modelo Naive para la clasificación Bayesiana, para cierto conjunto de datos de entrenamiento.

Primeramente, se obtienen las probabilidades a priori, .

***Salida de R:***

|  |  |
| --- | --- |
| **Probabilidades A-priori** | |
| **No** | **Si** |
| 0.6767 | 0.3233 |

Se observa que la probabilidad de que la persona no sobreviva es 0.6767, y que la probabilidad de sobrevivir es 0.3233.

Se obtienen las probabilidades condicionales, **,** por cada variable.

Se muestra la salida para la variable Class:

***Salida de R:***

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Variable: Class** | | | | |
|  | 1st | 2nd | 3rd | Tripulación |
| **No** | 0.0823 | 0.1142 | 0.3619 | 0.4416 |
| **Sí** | 0.2759 | 0.1757 | 0.2443 | 0.3040 |

De esta tabla se obtiene la siguiente información:

* La probabilidad de que una persona sea de la clase 1 dado que no sobrevivió es 0.0823.
* La probabilidad de que una persona sea de la clase 1 dado que sobrevivió es 0.2759.
* La probabilidad de que una persona sea de la clase 2 dado que no sobrevivió es 0.1142.
* La probabilidad de que una persona sea de la clase 2 dado que sobrevivió es 0.1757.
* La probabilidad de que una persona sea de la clase 3 dado que no sobrevivió es 0.3619.
* La probabilidad de que una persona sea de la clase 3 dado que sobrevivió es 0.2443.
* La probabilidad de que una persona sea de la Tripulación dado que no sobrevivió es 0.4416.
* La probabilidad de que una persona sea de la Tripulación dado que sobrevivió es 0.3040.

|  |  |  |
| --- | --- | --- |
| **Variable: Sex** | | |
|  | Hombre | Mujer |
| **No** | 0.9127 | 0.0873 |
| **Si** | 0.5167 | 0.4833 |

* La probabilidad de que una persona sea hombre dado que no sobrevivió es 0.9127.
* La probabilidad de que una persona sea mujer dado que no sobrevivió es 0.0873.
* La probabilidad de que una persona sea hombre dado que sobrevivió es 0.5167.
* La probabilidad de que una persona sea mujer dado que sobrevivió es 0.4833.

Se procede a validar los resultados, para ello haremos una predicción utilizando el conjunto de prueba.

La regla de clasificación del modelo se describe en la siguiente tabla:

***Salida de R:***

|  |  |  |
| --- | --- | --- |
| **Clase** | **Sexo** | **Predicción** |
| 1 | Hombre | **No** |
| 2 | Hombre | **No** |
| 3 | Hombre | **No** |
| Tripulación | Hombre | **No** |
| 1 | Mujer | **Si** |
| 2 | Mujer | **Si** |
| 3 | Mujer | **Si** |
| Tripulación | Mujer | **Si** |

De la tabla podemos ver que, sin importar la clase, el modelo clasificó a todas las mujeres como sobrevivientes, y a todos los hombres como que no lo hicieron.

***Salida de R:***

|  |  |  |  |
| --- | --- | --- | --- |
|  | | **Valores Reales** | |
|  |  | **No** | **Sí** |
| **Predicción** | **No** | **277** | **73** |
| **Sí** | **22** | **69** |
| **Exactitud: 78.46%**  **Sensibilidad: 0.4859**  **Especifidad: 0.9264**  **Precisión (**Pos Pred Value**): 0.7582**  **'Positive' Class : Si** | | | |

Se muestran los valores reales y los valores predichos, además el modelo tiene un 78.46% de exactitud.

De las 350 personas que no sobrevivieron de acuerdo con la predicción, 73 resultaron erróneos ya que sí sobrevivieron. Por otra parte, de las 91 personas que sobrevivieron de acuerdo con la predicción, 22 resultaron como no sobrevivientes.

De los 441 datos de prueba, 299 no sobrevivieron (casos “negativos”) y 142 sí sobrevivieron (casos “positivos”).

Sensibilidad:

Los “verdaderos positivos” son los que el modelo los clasificó como que si sobrevivieron y realmente lo hicieron que son 69. Por lo que

Especificidad:

Los “verdaderos negativos” son los que el modelo los clasificó como que no sobrevivieron y realmente no lo hicieron que son 277. Por lo que

Precisión:

Exactitud:

De los 441 registros, 277 y 69 fueron predichos de manera correcta por el modelo, por lo que:

**Sexo-Edad.**

Se ajusta el modelo Naive para la clasificación Bayesiana, para cierto conjunto de datos de entrenamiento.

Primeramente, se obtienen las probabilidades a priori, .

***Salida de R:***

|  |  |
| --- | --- |
| **Probabilidades A-priori** | |
| **No** | **Si** |
| 0.6727 | 0.3273 |

Se observa que la probabilidad de que la persona no sobreviva es 0.6727, y que la probabilidad de sobrevivir es 0.3273.

Se obtienen las probabilidades condicionales, **,** por cada variable.

Se muestra la salida para la variable Sexo:

***Salida de R:***

|  |  |  |
| --- | --- | --- |
| **Variable: Sex** | | |
|  | Hombre | Mujer |
| **No** | 0.9155 | 0.0845 |
| **Si** | 0.5190 | 0.4810 |

De la tabla anterior se recaba la siguiente información:

* La probabilidad de que una persona sea hombre dado que no sobrevivió es 0.9155.
* La probabilidad de que una persona sea mujer dado que no sobrevivió es 0.0845.
* La probabilidad de que una persona sea hombre dado que sobrevivió es 0.5190.
* La probabilidad de que una persona sea mujer dado que sobrevivió es 0.4810.

Para la variable Edad:

|  |  |  |
| --- | --- | --- |
| **Variable: Age** | | |
|  | Adulto | Niño |
| **No** | 0.9619 | 0.0381 |
| **Si** | 0.9184 | 0.0816 |

* La probabilidad de que una persona sea adulto dado que no sobrevivió es 0.9619.
* La probabilidad de que una persona sea niño dado que no sobrevivió es 0.0381.
* La probabilidad de que una persona sea adulto dado que sobrevivió es 0.9184.
* La probabilidad de que una persona sea niño dado que sobrevivió es 0.0816.

Se procede a validar los resultados, para ello haremos una predicción utilizando el conjunto de prueba.

La regla de clasificación del modelo se describe en la siguiente tabla:

***Salida de R:***

|  |  |  |
| --- | --- | --- |
| **Sexo** | **Edad** | **Predicción** |
| Hombre | Niño | **No** |
| Hombre | Adulto | **No** |
| Mujer | Niño | **Si** |
| Mujer | Adulto | **Si** |

De la tabla podemos ver que, el modelo no hace distinción de la edad del tripulante, es decir, sin importar la edad se clasificó a todas las mujeres como sobrevivientes, y a todos los hombres como que no lo hicieron.

***Salida de R:***

|  |  |  |  |
| --- | --- | --- | --- |
|  | | **Valores Reales** | |
|  |  | **No** | **Sí** |
| **Predicción** | **No** | **280** | **68** |
| **Sí** | **26** | **67** |
| **Exactitud: 78.68%**  **Sensibilidad: 0.4963**  **Especifidad: 0.9150**  **Precisión (**Pos Pred Value**): 0.7204**  **'Positive' Class : Si** | | | |

Se muestran los valores reales y los valores predichos, además el modelo tiene un 78.68% de exactitud.

De las 348 personas que no sobrevivieron de acuerdo con la predicción, 68 resultaron erróneos ya que sí sobrevivieron. Por otra parte, de las 93 personas que sobrevivieron de acuerdo con la predicción, 26 resultaron como no sobrevivientes.

De los 441 datos de prueba, 306 no sobrevivieron (casos “negativos”) y 135 sí sobrevivieron (casos “positivos”).

Los indicadores de exactitud para este modelo son los siguientes:

Sensibilidad:

Los “verdaderos positivos” son los que el modelo los clasificó como que si sobrevivieron y realmente lo hicieron que son 67. Por lo que

Especificidad:

Los “verdaderos negativos” son los que el modelo los clasificó como que no sobrevivieron y realmente no lo hicieron que son 280. Por lo que

Precisión:

Exactitud:

De los 441 registros, 280 y 67 fueron predichos de manera correcta por el modelo, por lo que:

**Modelo de todas las variables.**

Se ajusta el modelo Naive para la clasificación Bayesiana, tomando un conjunto de datos de entrenamiento.

Primeramente, se obtienen las probabilidades a priori, .

***Salida de R:***

|  |  |
| --- | --- |
| **Probabilidades A-priori** | |
| **No** | **Si** |
| 0.6795455 | 0.3204545 |

Se observa que la probabilidad de que la persona no sobreviva es 0.67954, y que la probabilidad de sobrevivir es 0.320454.

Se obtienen las probabilidades condicionales, **,** por cada variable.

Se muestra la salida para la variable Class:

***Salida de R:***

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Variable: Class** | | | | |
|  | 1st | 2nd | 3rd | Tripulación |
| **No** | 0.089464 | 0.112040 | 0.356187 | 0.442307 |
| **Sí** | 0.276595 | 0.161347 | 0.258865 | 0.303191 |

De esta tabla se obtiene la siguiente información:

* La probabilidad de que una persona sea de la clase 1 dado que no sobrevivió es 0.08946.
* La probabilidad de que una persona sea de la clase 1 dado que sobrevivió es 0.27565.
* La probabilidad de que una persona sea de la clase 2 dado que no sobrevivió es 0.1120.
* La probabilidad de que una persona sea de la clase 2 dado que sobrevivió es 0.16134.
* La probabilidad de que una persona sea de la clase 3 dado que no sobrevivió es 0.35618.
* La probabilidad de que una persona sea de la clase 3 dado que sobrevivió es 0.25886.
* La probabilidad de que una persona sea de la tripulación dado que no sobrevivió es 0.4423.
* La probabilidad de que una persona sea de la tripulación dado que sobrevivió es 0.30319.

Se muestra la salida para la variable Sex:

***Salida de R:***

|  |  |  |
| --- | --- | --- |
| **Variable: Sex** | | |
|  | Hombre | Mujer |
| **No** | 0.918896 | 0.081103 |
| **Si** | 0.533687 | 0.466312 |

* La probabilidad de que una persona sea hombre dado que no sobrevivió es 0.91889.
* La probabilidad de que una persona sea hombre dado que sobrevivió es 0.53368.
* La probabilidad de que una persona sea mujer dado que no sobrevivió es 0.0811.
* La probabilidad de que una persona sea mujer dado que sobrevivió es 0.46631.

Se muestra la salida para la variable Age:

***Salida de R:***

|  |  |  |
| --- | --- | --- |
| **Variable: Age** | | |
|  | Adulto | Niño |
| **No** | 0.966555 | 0.033444 |
| **Si** | 0.911347 | 0.088652 |

* La probabilidad de que una persona sea adulto dado que no sobrevivió es 0.96655.
* La probabilidad de que una persona sea adulto dado que sobrevivió es 0.91134.
* La probabilidad de que una persona sea niño dado que no sobrevivió es 0.033444.
* La probabilidad de que una persona sea niño dado que sobrevivió es 0.08865.

Posteriormente, se evalúa el modelo con los datos de prueba.

La regla de clasificación del modelo se describe en la siguiente tabla:

***Salida de R:***

|  |  |  |  |
| --- | --- | --- | --- |
| **Clase** | **Sexo** | **Age** | **Predicción** |
| 1 | Hombre | Adulto | **No** |
| 2 | Hombre | Adulto | **No** |
| 3 | Hombre | Adulto | **No** |
| Tripulación | Hombre | Adulto | **No** |
| 2 | Hombre | Niño | **Sí** |
| 3 | Hombre | Niño | **No** |
| 1 | Mujer | Adulto | **Si** |
| 2 | Mujer | Adulto | **Si** |
| 3 | Mujer | Adulto | **Si** |
| Tripulación | Mujer | Adulto | **Si** |
| 2 | Mujer | Niño | **Si** |
| 3 | Mujer | Niño | **Si** |

De la tabla podemos ver que, en general, sin importar la clase, el modelo clasificó a todas las mujeres como sobrevivientes, y a casi todos los hombres como que no lo hicieron. Solo los hombres de la clase 2 que eran niños fueron clasificados como sobrevivientes.

Se procede a validar los resultados, para ello haremos una predicción utilizando el conjunto de prueba.

***Salida de R:***

|  |  |  |  |
| --- | --- | --- | --- |
|  | | **Valores Reales** | |
|  |  | **No** | **Sí** |
| **Predicción** | **No** | **265** | **64** |
| **Sí** | **29** | **83** |
| **Exactitud: 78.91%**  **Sensibilidad: 0.5646**  **Especifidad: 0.9014**  **Precisión (**Pos Pred Value**): 0.7411**  **'Positive' Class : Si** | | | |

Se muestran los valores reales y los valores predichos, además el modelo tiene un 78.91% de exactitud.

El modelo predijo que 329 personas no sobrevivieron. De estas predicciones, 64 resultaron erróneas. Por otra parte, de las 112 personas que sobrevivieron de acuerdo con la predicción, 29 resultaron como no sobrevivientes.

De los 441 datos de prueba, 294 no sobrevivieron (casos “negativos”) y 147 sí sobrevivieron (casos “positivos”).

Sensibilidad:

Los “verdaderos positivos” son los que el modelo los clasificó como que si sobrevivieron y realmente lo hicieron que son 83. Por lo que

Especificidad:

Los “verdaderos negativos” son los que el modelo los clasificó como que no sobrevivieron y realmente no lo hicieron que son 265. Por lo que

Precisión:

Exactitud:

De los 441 registros, 265 y 83 fueron predichos de manera correcta por el modelo, por lo que:

La tabla siguiente resume los indicadores de exactitud de los tres modelos estudiados:

|  |  |  |  |
| --- | --- | --- | --- |
|  | Modelo 1 (Clase y Sexo) | Modelo 2 (Sexo y Edad) | Modelo 3 (Clase, Sexo y Edad) |
| Exactitud | 0.7845 | 0.7868 | 0.7891 |
| Sensibilidad | 0.4859 | 0.4963 | 0.5646 |
| Especificidad | 0.9264 | 0.9150 | 0.9014 |
| Precisión | 0.7582 | 0.7204 | 0.7411 |

Podemos observar que en Exactitud y Sensibilidad el modelo 3, aquel que considera todas las variables, es quien tiene los indicadores más cercanos a uno, mientras que en Especificidad y Precisión es el modelo 1, aquel que considera las variables clase y sexo, quien tiene los indicadores más altos. Adicionalmente, de los tres modelos, en cuanto a la predicción, fue fácil notar que la variable Sexo es la que mayormente nos predijo si un tripulante sobrevivió o no. Basándonos en el indicador de Exactitud **escogeríamos el modelo 3** debido a que es el que tiene el indicador más alto.