Proyecto final - Reconocimiento estadístico de patrones

Erick Salvador Alvarez Valencia

CIMAT A.C.,
erick.alvarez@cimat.mx

Resumen En el presente reporte se hablará sobre la resolución de los ejercicios del proyecto final de reconocimiento de patrones, de la misma manera se presentará el código de los mismos y los resultados que fueron obtenidos en los problemas que lo requieren. El código con el que se trabajó para los ejercicios se incluirá en los archivos del proyecto.

1. Problema 2

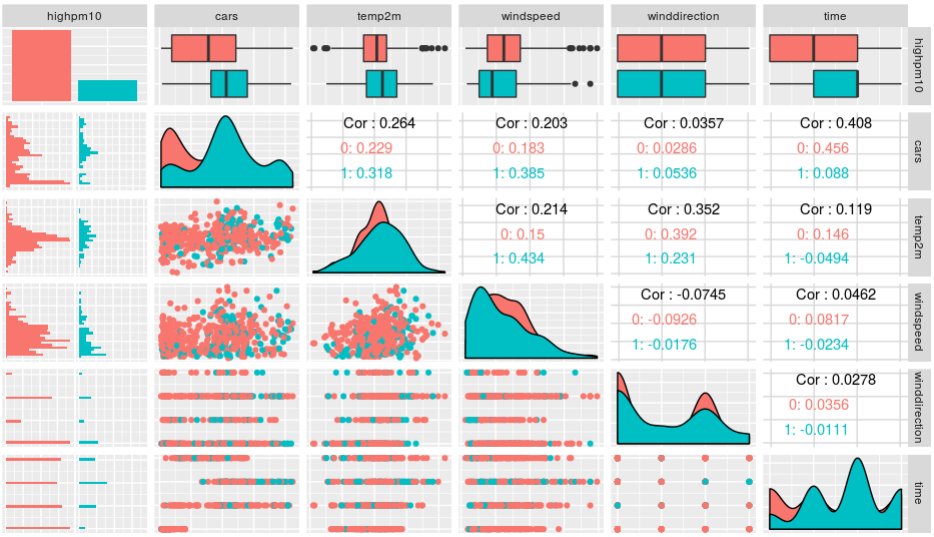
Este ejercicio es sobre modelos de regresión logística tomado de *http://www.maths.lth.se/matstat/kurser/masm22/*.
The data is a random subsample of 500 observations from a data set that originates in a study where air pollution at a road is related to traffic volume and meteorological variables, collected by the Norwegian Public Roads Administration. The data is hourly measurements at Alnabruin Oslo, Norway, between October 2001 and August 2003. In order to get rid of the strong correlation between successive measurements a random sample of the original, larger, data has been taken. The objective is to model the probability that the concentration of atmospheric particles with a diameter between 2.5 and 10 μ_m, PM 10, exceeds the limit 50 $\mu g/m^3$. This is the Swedish limit for the daily average but since we do not have access to the daily averages we will compare our hourly values to this limit.
Ajusta y evalua modelos de regresión logística para predecir si se rebasa o no el limite de 50 $\mu g/m^3$ usando el tiempo, dirección de viento, la temperatura a 2 metros sobre el suelo y el número de coche por hora. Compara el desempeño con unas redes neuronales adecuadas.

Variables:

highpm10	the concentration of PM_{10} particles (categorical), 0 = $PM_{10} \leq 50\,\mu g/m^3$, 1 = $PM_{10} > 50\,\mu g/m^3$
cars	the number of cars per hour
temp2m	temperature 2 meters above ground (degree C)
windspeed	wind speed (meters/second)
winddirection	wind direction (categorical), 1 = NE, 2 = SE, 3 = SW, 4 = NW
time	time of day (categorical), 1 = 01–06, 2 = 06–12, 3 = 12–18, 4 = 18–24

(a)

Solución : Lo primero que se hizo fue analizar las variables para poder encontrar algunas dependencias entre ellas y lo que se obtuvo fue el siguiente:



(b) Figura 1. Pairsplot de las variables del problema.

En la Figura 1. se muestran las correlaciones de las variables del problema, en ella podemos apreciar que algunas variables están fuertemente relacionadas, tal como los autos con el tiempo, la temperatura que está dos metros sobre el suelo con la dirección del viento, entre otras. Además podemos notar que existen más datos que contienen información de que la probabilidad de no estar contaminado sea más alta y para la variable de tiempo vemos que la contaminación ocurre en cuatro etapas distintas pero hay una en la que se conglomera más. Dichas correlaciones nos pueden servir más adelante para proponer algunos modelos de

regresión logística.

Al trabajar con regresión logística, el primer modelo que se probó fue el saturado, el cual contiene todos los predictores, esto se hizo para ver cuáles de ellos tenian más importancia a la hora de hacer predicciones. Para hacer esto se usó la librería *caret* con el método glm y la familia binomial, los datos fueron divididos en los conjuntos de prueba y entrenamiento aleatoriamente, esto para posteriormente probar el poder del predictor generado. A continuación se muestran los resultados obtenidos con este modelo.

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.2525  -0.7850  -0.5327  -0.3478   2.2649

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)   -1.825304   0.565944  -3.225 0.001259 **
cars           0.000445   0.000128   3.477 0.000507 ***
temp2m         0.026349   0.023168   1.137 0.255420
windspeed     -0.228739   0.086411  -2.647 0.008119 **
winddirection -0.137971   0.146104  -0.944 0.344999
time           0.269766   0.133691   2.018 0.043609 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 373.83  on 349  degrees of freedom
Residual deviance: 343.94  on 344  degrees of freedom
AIC: 355.94

Number of Fisher Scoring iterations: 4
```

(c) Figura 2. Resumen del modelo saturado.

Podemos ver en el resumen del modelo varias características mostradas de los predictores, una de ellas es la importancia que obtuvieron al hacer el entrenamiento. La variables cars, windspeed y time generaron mucha importancia a la hora de realizar las predicciones, esto lo confirma el P value asociado a las mismas, donde si tomamos un nivel de significancia del 5 % se rechaza la hipótesis nula que asume que estos predictores no son relevantes para el modelo. Posteriormente se hizo un test de predicción con este modelo entrenado y a continuación se muestran los resultados del mismo.

```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 113  33
         1   2   2

               Accuracy : 0.7667
                 95% CI : (0.6907, 0.8318)
    No Information Rate : 0.7667
    P-Value [Acc > NIR] : 0.5452

                  Kappa : 0.0575
 Mcnemar's Test P-Value : 3.959e-07

            Sensitivity : 0.98261
            Specificity : 0.05714
         Pos Pred Value : 0.77397
         Neg Pred Value : 0.50000
             Prevalence : 0.76667
         Detection Rate : 0.75333
   Detection Prevalence : 0.97333
      Balanced Accuracy : 0.51988

       'Positive' Class : 0
```

(d) Figura 3. Resultados de las predicciones realizadas por el modelo saturado.

Podemos ver en la Figura anterior que este modelo no obtuvo tan buenos resultados a la hora de predecir los datos del conjunto de prueba, obteniendo una precisión del 76 % de datos bien clasificados. Lo siguiente que se hizo fue proponer dos modelos basados en los parámetros más importantes reportados del modelo saturado. El primero de ellos incluye los predictores: $cars$, $winddirection$ y $temp2$, el segundo sólamente contiene $cars$ y $winddirection$. Dichas variables fueron las más significativas en el modelo anterior. A continuación se muestran los resultados obtenidos con los nuevos modelos.

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.1802  -0.7612  -0.5674  -0.4818   2.1657

Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)   -1.9683512  0.3860171  -5.099 3.41e-07 ***
cars           0.0004470  0.0001176   3.803 0.000143 ***
winddirection -0.0493533  0.1416189  -0.348 0.727470
temp2m         0.0128014  0.0224456   0.570 0.568455
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 373.83  on 349  degrees of freedom
Residual deviance: 356.07  on 346  degrees of freedom
AIC: 364.07

Number of Fisher Scoring iterations: 4
```

(e) Figura 4. Resumen del modelo 2.

```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 115  35
         1   0   0

               Accuracy : 0.7667
                 95% CI : (0.6907, 0.8318)
    No Information Rate : 0.7667
    P-Value [Acc > NIR] : 0.5452

                  Kappa : 0
 Mcnemar's Test P-Value : 9.081e-09

            Sensitivity : 1.0000
            Specificity : 0.0000
         Pos Pred Value : 0.7667
         Neg Pred Value :    NaN
             Prevalence : 0.7667
         Detection Rate : 0.7667
   Detection Prevalence : 1.0000
      Balanced Accuracy : 0.5000

       'Positive' Class : 0
```

(f) Figura 5. Resultados de las predicciones realizadas por el modelo 2.

Figura 1

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.1218  -0.7641  -0.5617  -0.4926   2.0790

Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)   -2.0445358  0.3635732  -5.623 1.87e-08 ***
cars           0.0004636  0.0001139   4.069 4.72e-05 ***
winddirection -0.0176678  0.1298638  -0.136    0.892
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 373.83  on 349  degrees of freedom
Residual deviance: 356.39  on 347  degrees of freedom
AIC: 362.39

Number of Fisher Scoring iterations: 4
```

(a) Figura 6. Resumen del modelo 3.

```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 115  35
         1   0   0

               Accuracy : 0.7667
                 95% CI : (0.6907, 0.8318)
    No Information Rate : 0.7667
    P-Value [Acc > NIR] : 0.5452

                  Kappa : 0
 Mcnemar's Test P-Value : 9.081e-09

            Sensitivity : 1.0000
            Specificity : 0.0000
         Pos Pred Value : 0.7667
         Neg Pred Value :    NaN
             Prevalence : 0.7667
         Detection Rate : 0.7667
   Detection Prevalence : 1.0000
      Balanced Accuracy : 0.5000

       'Positive' Class : 0
```

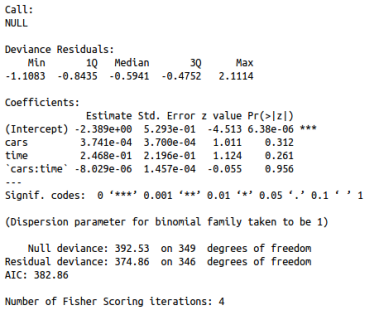
(b) Figura 7. Resultados de las predicciones realizadas por el modelo 3.

Figura 2

En las 4 Figuras anteriores podemos apreciar los resultados ofrecidos por los dos modelos propuestos. Primeramente en el resumen de ambos notamos que las variables más importantes de los mismos son el intercepto (la variable libre) y los carros, pero lo más importante es que al trabajarlos con los datos de prueba, ambos modelos dieron la misma precisión que el modelo saturado, es decir, no se obtuvo una mejora significativa al remover las variables menos importantes del saturado.
Las siguientes dos propuestas fueron basadas en las correlaciones de las variables mostradas al principio del ejercicio, en donde se pudo apreciar que las variables cars y tiempo así como temp2 y winddirection tienen un alto grado de correlación, por lo cual se propusieron los siguientes modelos:

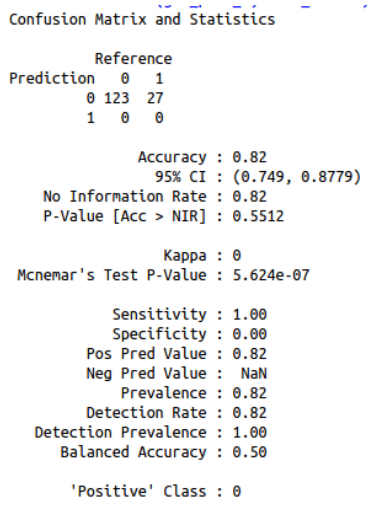
1. Y cars * time
2. Y cars * windspeed + temp2m * winddirection

A continuación se muestran los resultados obtenidos con los modelos previamente propuestos.



```
Call:
NULL

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.1083  -0.8435  -0.5941  -0.4752   2.1114

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.389e+00  5.293e-01  -4.513 6.38e-06 ***
cars         3.741e-04  3.700e-04   1.011    0.312
time         2.468e-01  2.196e-01   1.124    0.261
`cars:time` -8.029e-06  1.457e-04  -0.055    0.956
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 392.53  on 349  degrees of freedom
Residual deviance: 374.86  on 346  degrees of freedom
AIC: 382.86

Number of Fisher Scoring iterations: 4
```

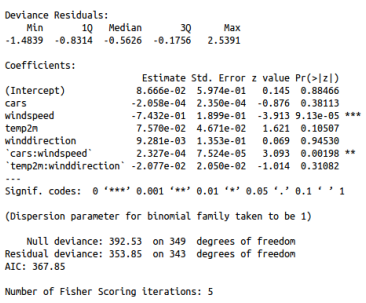
(a) Figura 8. Resumen del modelo 4.



```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 123  27
         1   0   0

               Accuracy : 0.82
                 95% CI : (0.749, 0.8779)
    No Information Rate : 0.82
    P-Value [Acc > NIR] : 0.5512

                  Kappa : 0
 Mcnemar's Test P-Value : 5.624e-07

            Sensitivity : 1.00
            Specificity : 0.00
         Pos Pred Value : 0.82
         Neg Pred Value :  NaN
             Prevalence : 0.82
         Detection Rate : 0.82
   Detection Prevalence : 1.00
      Balanced Accuracy : 0.50

       'Positive' Class : 0
```

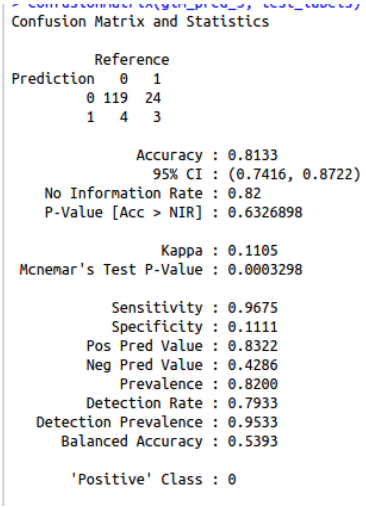
(b) Figura 9. Resultados de las predicciones realizadas por el modelo 4.

Figura 3



```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.4839  -0.8314  -0.5626  -0.1756   2.5391

Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)             8.666e-02  5.974e-01   0.145  0.88466
cars                   -2.058e-04  2.350e-04  -0.876  0.38113
windspeed              -7.432e-01  1.899e-01  -3.913 9.13e-05 ***
temp2m                  7.570e-02  4.671e-02   1.621  0.10507
winddirection           9.281e-03  1.353e-01   0.069  0.94530
`cars:windspeed`        2.327e-04  7.524e-05   3.093  0.00198 **
`temp2m:winddirection` -2.077e-02  2.050e-02  -1.014  0.31082
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 392.53  on 349  degrees of freedom
Residual deviance: 353.85  on 343  degrees of freedom
AIC: 367.85

Number of Fisher Scoring iterations: 5
```

(a) Figura 10. Resumen del modelo 5.



```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 119  24
         1   4   3

               Accuracy : 0.8133
                 95% CI : (0.7416, 0.8722)
    No Information Rate : 0.82
    P-Value [Acc > NIR] : 0.6326898

                  Kappa : 0.1105
 Mcnemar's Test P-Value : 0.0003298

            Sensitivity : 0.9675
            Specificity : 0.1111
         Pos Pred Value : 0.8322
         Neg Pred Value : 0.4286
             Prevalence : 0.8200
         Detection Rate : 0.7933
   Detection Prevalence : 0.9533
      Balanced Accuracy : 0.5393

       'Positive' Class : 0
```

(b) Figura 11. Resultados de las predicciones realizadas por el modelo 5.

Figura 4

En las 4 Figuras anteriores podemos ver los resultados de los modelos descritos en donde el primero es el producto de las variables cars y time y el segundo es el producto de las variables cars y windspeed sumado el producto de temp2m y winddirection. En el primero de esos modelos podemos ver que el nivel de significancia de las variables sólamente se marca con el intercepto aunque si nos fijamos en los resultados de las predicciones hechas con el conjunto de datos de prueba podemos ver que el modelo 4 obtuvo una precisión del 82 % una gran mejora con respecto a los modelos anteriores.
Por otra parte podemos ver que en resumen del modelo 5 la variable winddirection y el conjunto de cars con windspeed se marcaron como las más significativas al momento de entrenar, y finalmente el modelo obtuvo un 81 % de precisión con respecto a los datos de prueba, un poco más bajo que el modelo anterior, pero de la misma forma mejoró a los 3 modelos anteriores. Una cosa muy importante que podemos destacar del último modelo es que este fue el único en predecir datos de la clase 1, es decir, cuando hay mucha probabilidad de contaminación, esto lo vemos en la matriz de confusión en la primer parte donde el modelo predijo tres verdaderos positivos y cuatro falsos positivos, ningún otro modelo tuvo esta característica.
Para concluir esta sección de regresión logística, se trabajó con el modelo sugerido por el ejercicio el cual contiene las siguientes variables predictoras: *time*, *winddirection*, *temp*2*m* y *cars*. Los resultados obtenidos fueron los siguientes:

```
> summary(gln_model_ex)

Call:
NULL

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.2136  -0.7674  -0.5144  -0.3803   2.3333

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)   -2.8574895  0.5463726  -5.230 1.70e-07 ***
time           0.1676740  0.1464051   1.145    0.252
winddirection  0.0153300  0.1432660   0.107    0.915
temp2m        -0.0071670  0.0244272  -0.293    0.769
cars           0.0005846  0.0001309   4.466 7.97e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 366.29  on 349  degrees of freedom
Residual deviance: 336.02  on 345  degrees of freedom
AIC: 346.02

Number of Fisher Scoring iterations: 4
```

(a) Figura 12. Resumen del modelo 5.

```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 111  38
         1   1   0

               Accuracy : 0.74
                 95% CI : (0.6621, 0.8081)
    No Information Rate : 0.7467
    P-Value [Acc > NIR] : 0.6163

                  Kappa : -0.0132
 Mcnemar's Test P-Value : 8.185e-09

            Sensitivity : 0.9911
            Specificity : 0.0000
         Pos Pred Value : 0.7450
         Neg Pred Value : 0.0000
             Prevalence : 0.7467
         Detection Rate : 0.7400
   Detection Prevalence : 0.9933
      Balanced Accuracy : 0.4955

       'Positive' Class : 0
```

(b) Figura 13. Resultados de las predicciones realizadas por el modelo 5.

Figura 5

En el modelo anterior se puede ver que para el entrenamiento la variable más relevante fue *cars* y en la parte de las predicciones con los datos de prueba se obtuvo un porcentaje del 74 % de aciertos, un poco menor que el modelo saturado, por lo cual vemos que no fue tan eficiente este modelo con las predicciones.

Posteriormente se generaron algunos modelos con redes neuronales que trabajaran con el mismo conjunto de datos, dichos modelos se trabajaron con diferentes cantidades de capas ocultas, así como diferentes valores de *learning rate* (tamaño de paso) en el algoritmo de optimización. El enfoque con dichos modelos correspondió mas a la fórmula indicada en la red. Así como en los modelos de regresión, primero se trabajó con el modelo saturado de la red. Para dicho modelo estos fueron los resultados obtenidos.

```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 111  27
         1   6   6

               Accuracy : 0.78
                 95% CI : (0.7051, 0.8435)
    No Information Rate : 0.78
    P-Value [Acc > NIR] : 0.5464760

                  Kappa : 0.1692
 Mcnemar's Test P-Value : 0.0004985

            Sensitivity : 0.9487
            Specificity : 0.1818
         Pos Pred Value : 0.8043
         Neg Pred Value : 0.5000
             Prevalence : 0.7800
         Detection Rate : 0.7400
   Detection Prevalence : 0.9200
      Balanced Accuracy : 0.5653

       'Positive' Class : 0
```

(a) Figura 13. Resultados de las predicciones realizadas por el modelo saturado de NNET.

Podemos ver en la Figura anterior que dicho modelo generó un 78 % de precisión con los datos de prueba, curiosamente sólo clasificó tanto falsos negativos como verdaderos negativos, todos los datos relacionados a la clase 1 los clasificó mal.
Posteriormente se generó un segundo modelo basado en los predictores más importantes del modelo saturado de la regresión logística, los cuales fueron: $cars$, $winddirection$ y $temp2m$ esto para ver qué tan bien trabajaban con la red. Esto fue el resultado obtenido.

```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 117  33
         1   0   0

               Accuracy : 0.78
                 95% CI : (0.7051, 0.8435)
    No Information Rate : 0.78
    P-Value [Acc > NIR] : 0.5465

                  Kappa : 0
 Mcnemar's Test P-Value : 2.54e-08

            Sensitivity : 1.00
            Specificity : 0.00
         Pos Pred Value : 0.78
         Neg Pred Value :  NaN
             Prevalence : 0.78
         Detection Rate : 0.78
   Detection Prevalence : 1.00
      Balanced Accuracy : 0.50

       'Positive' Class : 0
```

(b) Figura 14. Resultados de las predicciones realizadas por el modelo 2 de NNET.

Se puede observar que se obtuvo el mismo resultado que con el modelo anterior, en este caso no trabajaron bien las variables predictoras elegidas, de igual manera vemos que no se clasificó ningun dato de la clase 1.
Finalmente se propuso un modelo sin la variable *cars* ya que se cree que esta variable es muy influyente en la predicción de los datos, a continuación se muestran los resultados obtenidos.

```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 110  27
         1   7   6

               Accuracy : 0.7733
                 95% CI : (0.6979, 0.8376)
    No Information Rate : 0.78
    P-Value [Acc > NIR] : 0.62259

                  Kappa : 0.1559
 Mcnemar's Test P-Value : 0.00112

            Sensitivity : 0.9402
            Specificity : 0.1818
         Pos Pred Value : 0.8029
         Neg Pred Value : 0.4615
             Prevalence : 0.7800
         Detection Rate : 0.7333
   Detection Prevalence : 0.9133
      Balanced Accuracy : 0.5610

       'Positive' Class : 0
```

(c) Figura 15. Resultados de las predicciones realizadas por el modelo 3 de NNET.

Podemos ver que este modelo obtuvo un porcentaje de precisión distinto a los anteriores con los datos de prueba, se consiguió obtener un 77 % de precisión el cual es un poco menor del conseguido por los dos modelos anteriores, aunque en este modelo se puede ver que consiguió clasificar datos de la clase 1, por lo cual se prefiere este modelo a los dos anteriores.
Finalmente se concluye que con los modelos de regresión logística se obtuvo mejores resultados que los de redes neuronales, además se prefiere el penúltimo modelo de regresión logística ya que aunque tuvo una precisión un poco más baja que su antecesor este pudo ser capaz de predecir datos de la clase 1.

2. Problema 3

Construye, evalua y compara clasificadores basados en regresión logistica para los datos SPAM de la tarea anterior. Compara su poder predictivo con algunos otros clasificadores que vimos en clase.

Solución : Lo primero que se hizo fue leer los datos y partirlos en dos conjuntos, uno de entrenamiento con el 70 % de las entradas y el de prueba con el 30 % restante, la partición se hizo de forma alteatoria, de tal forma que ambos conjuntos contuvieran datos pertenecientes a las dos clases y esto evitara que los clasificadores se fueran a sobreajustar. Posteriormente se usó la librería *caret* para usar los clasificadores con respecto a los datos de SPAM, el uso de esta librería fue por el hecho de que la misma contiene un amplio conjunto de clasificadores, entre los que ya hemos visto en las clases, además de que prove

varias funciones interesantes, como la validación cruzada, la cual sirve bastante para evitar un sobreajuste en los clasificadores. Para poder hacer una comparación justa con respecto al mismo conjunto de datos, con cada clasificador se usó la misma fórmula, el mismo número de iteraciones, así como una selección de parámetros amplia y se le daba la opción al método train para conservar el conjunto de parámetros que haya dado mejores resultados en la fase de entrenamiento. A continuación se muestran los clasificadores usados así como los parámetros que se incluyeron a cada uno:

1. **Regresión logística** : Familia: binomial.
2. **Boosting** : Número máximo de ramas en los árboles: De 1 a 3.
3. **SVM** : Kernel: lineal, factor de penalización (gamma): {0.01, 0.1, 0.5, 1.5, 2.5, 5, 10}.
4. **KNN** : Número de vecinos: {2, 3, 4, 5}.
5. **NNet** : Número de capas ocultas: {1, 2, 3}, parámetro de aprendizaje: {0.001, 0.1, 1.0, 2.0, 5.0}.

Ahora se mostrarán los resultados obtenidos con los diferentes clasificadores usados:

1. **Regresión logística** : Para este método de clasificación, se obtuvieron los siguientes resultados con el conjunto de datos de prueba.

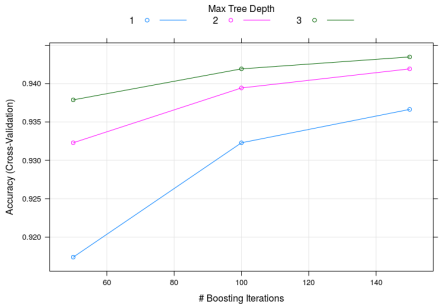
```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 781  64
         1  40 496

               Accuracy : 0.9247
                 95% CI : (0.9095, 0.9381)
    No Information Rate : 0.5945
    P-Value [Acc > NIR] : < 2e-16

                  Kappa : 0.8427
 Mcnemar's Test P-Value : 0.02411

            Sensitivity : 0.9513
            Specificity : 0.8857
         Pos Pred Value : 0.9243
         Neg Pred Value : 0.9254
             Prevalence : 0.5945
         Detection Rate : 0.5655
   Detection Prevalence : 0.6119
      Balanced Accuracy : 0.9185

       'Positive' Class : 0
```

(d) Figura 16. Resultados de las predicciones realizadas por el modelo saturado.

Se puede ver que este método de clasificación para el conjunto de datos de prueba obtuvo una precisión del 92 %. Además en la primer parte de

la matriz de confusión podemos notar que el clasificador predijo el mismo porcentaje de verdaderos positivos que verdaderos negativos, lo que significa que no se sobreajustó a los datos de cierta clase.

2. **Boosting** : Para esta parte se usó un clasificador de boosting construido con árboles de decisión con diferentes profundidades (se escogió un rango de profundidades de 1 a 3). Estos fueron los resultados obtenidos:



(e) Figura 17. Gráfica del modelo de boosting.

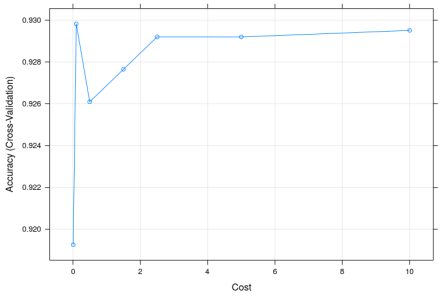
```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 794  53
         1  27 507

               Accuracy : 0.9421
                 95% CI : (0.9284, 0.9538)
    No Information Rate : 0.5945
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.879
 Mcnemar's Test P-Value : 0.005189

            Sensitivity : 0.9671
            Specificity : 0.9054
         Pos Pred Value : 0.9374
         Neg Pred Value : 0.9494
             Prevalence : 0.5945
         Detection Rate : 0.5749
   Detection Prevalence : 0.6133
      Balanced Accuracy : 0.9362

       'Positive' Class : 0
```

(f) Figura 18. Resultados de las predicciones realizadas por el modelo de boosting.

Figura 6

En la Figura 17 podemos ver los resultados obtenidos con los diferentes niveles de profundidad de los árboles y lo que notamos es que con 3 niveles de profundidad se obtuvo el porcentaje más alto de predicción el cual fue un poco más de 94 %. Posteriormente en la Figura 13 se muestran los resultados de la predicción con el conjunto de prueba, en donde vemos que el clasificador de boosting obtuvo de igual manera un 94 % de aciertos en sus predicciones, un poco más que el modelo de regresión logística.

3. **Máquina de soporte vectorial** : Para esta parte se usó un clasificador de SVM con diferentes valores de penalización en el parámetro γ además de usar un kernel lineal. Estos fueron los resultados obtenidos:



(a) Figura 20. Gráfica del modelo de SVM.

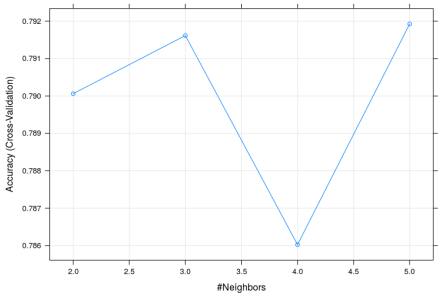
```
          Reference
Prediction   0   1
         0 785  64
         1  36 496

               Accuracy : 0.9276
                 95% CI : (0.9126, 0.9407)
    No Information Rate : 0.5945
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.8486
 Mcnemar's Test P-Value : 0.006934

            Sensitivity : 0.9562
            Specificity : 0.8857
         Pos Pred Value : 0.9246
         Neg Pred Value : 0.9323
             Prevalence : 0.5945
         Detection Rate : 0.5684
   Detection Prevalence : 0.6148
      Balanced Accuracy : 0.9209

       'Positive' Class : 0
```

(b) Figura 21. Resultados de las predicciones realizadas por el modelo de SVM.

Figura 7

En la Figura 20 se pueden ver los resultados del entrenamiento de la SVM con los diferentes valores de penalización, y lo que podemos notar es que para el valor $\gamma = 0{,}01$ se obtuvieron los mejores resultados con el conjunto de entrenamiento, posteriormente hubo un decremento en la calidad y finalmente entre más aumentaba el parámetro de penalización más aumentaba la calidad de las predicciones. Posteriormente en la Figura 15 se nos muestran los resultados de las predicciones con el conjunto de prueba, donde el clasificador obtuvo un 92 % de precisión con dicho conjunto, un nivel de predicción muy bueno aunque igualó a la regresión logística.

4. **KNN** : Para esta parte se usó un clasificador de KNN con diferentes números de vecinos (2, 3, 4 y 5). Estos fueron los resultados obtenidos:



(a) Figura 22. Gráfica del modelo de KNN.

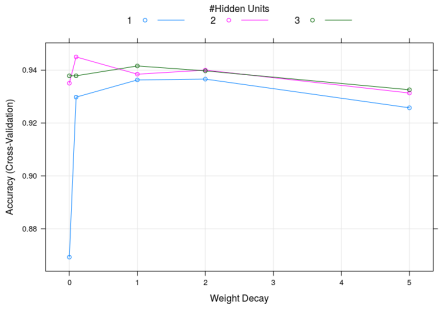
```
          Reference
Prediction   0   1
         0 690 157
         1 131 403

               Accuracy : 0.7915
                 95% CI : (0.7691, 0.8126)
    No Information Rate : 0.5945
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.5642
 Mcnemar's Test P-Value : 0.1407

            Sensitivity : 0.8404
            Specificity : 0.7196
         Pos Pred Value : 0.8146
         Neg Pred Value : 0.7547
             Prevalence : 0.5945
         Detection Rate : 0.4996
   Detection Prevalence : 0.6133
      Balanced Accuracy : 0.7800

       'Positive' Class : 0
```

(b) Figura 23. Resultados de las predicciones realizadas por el modelo de KNN.

Figura 8

Podemos ver en la Figura 22 los resultados del entrenamiento del KNN con cada parámetro de número de vecinos indicado, este obtuvo su mejor valor con 5 vecinos para el conjunto de entrenamiento el cual fue de 79 % de precisión, y posteriormente cuando se probó con el conjunto de prueba se logró de igual manera un 79 % de aciertos, en donde logró identificar mejor verdaderos positivos que verdaderos negativos. Hasta el momento este es el clasificador que peores resultados a ofrecido con el mismo conjunto de datos.

5. **NNET** : Para esta parte se usó un clasificador basado en redes neuronales con diferentes números de capas ocultas (1, 2 y 3) además diferentes valores de learning rate (0.001, 0.1, 1.0, 2.0 y 5.0). Estos fueron los resultados obtenidos:



(a) Figura 24. Gráfica del modelo de NNET.

```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 781  55
         1  40 505

               Accuracy : 0.9312
                 95% CI : (0.9166, 0.944)
    No Information Rate : 0.5945
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.8567
 Mcnemar's Test P-Value : 0.1509

            Sensitivity : 0.9513
            Specificity : 0.9018
         Pos Pred Value : 0.9342
         Neg Pred Value : 0.9266
             Prevalence : 0.5945
         Detection Rate : 0.5655
   Detection Prevalence : 0.6054
      Balanced Accuracy : 0.9265

       'Positive' Class : 0
```

(b) Figura 25. Resultados de las predicciones realizadas por el modelo de NNET.

Figura 9

Para este último clasificador trabajado podemos ver en la Figura 24 los resultados del entrenamiento para el diferente número de capas ocultas y cada una con el diferente número de learning rate, lo que se puede notar es que el mejor ajuste se logró con 3 capas ocultas y un learning rate 5, finalmente se ve en los resultados de predicción que se obtuvo un total de 93 % de predicción con este método.

Podemos concluir que con respecto a las pruebas realizadas con los diferentes clasificadores el que dió un mejor rendimiento fue la máquina de soporte vectorial con un parámetro de penalización de 0.01, siguiendo la red neuronal con 3 capas ocultas con un learning de 5.

3. Ejercicio 4

Los siguientes datos son de un estudio sobre la relación entre tomar la medicina Azidotimidina(AZT), la raza de un paciente y mostrar síntomas de sida (si o no).

Race	AZT Use	Symptoms	
		Yes	No
White	Yes	14	93
	No	32	81
Black	Yes	11	52
	No	12	43

Source: New York Times, Feb. 15, 1991.

(a)

Busca un modelo de regresión logística adecuada que relaciona la probabilidad de mostrar los síntomas de sida con tomar AZT y la raza. ¿Cuáles variables son influyentes?

Solución : Primero se formó el archivo con los datos, representando con valores binarios las variables AZT Use y Race, posteriormente se generaron varios modelos de regresión logística por proporciones los cuales se muestran a continuación, además de que por cada modelo se dará un análisis y finalmente las conclusiones:

1. **Modelo saturado** : Primero se trató el modelo más clásico que es el saturado, para dicho modelo este es el resumen obtenido:

```
Call:
glm(formula = cbind(Y, N) ~ Race + AZT, family = binomial(logit),
    data = data)

Deviance Residuals:
      1       2       3       4
-0.5547  0.4253  0.7035 -0.6326

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.07357    0.26294  -4.083 4.45e-05 ***
Race         0.05548    0.28861   0.192  0.84755
AZT         -0.71946    0.27898  -2.579  0.00991 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 8.3499  on 3  degrees of freedom
Residual deviance: 1.3835  on 1  degrees of freedom
AIC: 24.86

Number of Fisher Scoring iterations: 4
```

(b) Figura 26. Resumen del modelo 1.

Para este modelo podemos ver que el P value de la variable AZT es bastante pequeño y por lo tanto se rechaza la hipótesis nula que asume este dato como no significativo en el modelo, en cambio la variable de raza no generó tanto impacto en el entrenamiento con respecto al modelo, además se aplicó una prueba pseudo-R2 al mismo modelo la cual es un indicador para ver qué tan bien este se ajustó al conjunto de datos, y de entre los datos obtenidos el coeficiente de McFadden fue de 0.26, dicho coeficiente se encarga de evaluar la calidad del modelo generado con respecto a las desviaciones nula y residual, esto en una escala de 0 a 1, por lo que se podría decir que este modelo no es tan bueno.

2. **Modelo con Race como única variable predictora** : Este modelo sólamente usó a la variable de raza como la predictora, a continuación se muestra el resumen obtenido:

```
Call:
glm(formula = cbind(Y, N) ~ Race, family = binomial(logit), data = data)

Deviance Residuals:
      1       2       3       4
-2.1028  1.8650 -0.4126  0.4294

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.41838    0.23239  -6.103 1.04e-09 ***
Race         0.08797    0.28547   0.308    0.758
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 8.3499  on 3  degrees of freedom
Residual deviance: 8.2544  on 2  degrees of freedom
AIC: 29.731

Number of Fisher Scoring iterations: 4
```

(c) Figura 27. Resumen del modelo 2.

Para este modelo podemos ver que el P value de la variable Race no fue tan pequeño por lo que se acepta la hipótesis nula de que esta variable no es tan relevante para predecir la información, la única variable relevante fue el intercepto. De la misma forma se aplicó la prueba R2 a este modelo y el coeficiente de McFadden fue de 0.00369, una cantidad bastante baja lo cual nos indica que el modelo es bastante malo con respecto a las desviaciones nula y estándar.

3. **Modelo con AZT como única variable predictora** : Este modelo sólamente usó a la variable de AZT (consumo del medicamento) como la predictora, a continuación se muestra el resumen obtenido:

```
Call:
glm(formula = cbind(Y, N) ~ AZT, family = binomial(logit), data = data)

Deviance Residuals:
      1        2        3        4
-0.4813   0.5102   0.6026  -0.7521

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -1.0361     0.1755  -5.904 3.54e-09 ***
AZT          -0.7218     0.2787  -2.590  0.00961 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 8.3499  on 3  degrees of freedom
Residual deviance: 1.4206  on 2  degrees of freedom
AIC: 22.897

Number of Fisher Scoring iterations: 4
```

(d) Figura 28. Resumen del modelo 3.

Podemos notar que para este modelo, la única variable predictora, la cual es AZT resultó de cierta manera significativa en el entrenamiento, lo mismo pasó con el modelo saturado. El coeficiente de McFadden para este modelo fue de 0.323.

4. **Modelo multiplicativo con los dos predictores** : El último modelo analizado fue el basado en la multiplicación de los dos predictores, la raza y el AZT para el cual se obtuvieron los siguientes resultados:

```
Call:
glm(formula = cbind(Y, N) ~ Race * AZT, family = binomial(logit),
    data = data)

Deviance Residuals:
[1]  0  0  0  0

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -1.2763     0.3265  -3.909 9.26e-05 ***
Race          0.3476     0.3875   0.897    0.370
AZT          -0.2771     0.4655  -0.595    0.552
Race:AZT     -0.6878     0.5852  -1.175    0.240
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 8.3499e+00  on 3  degrees of freedom
Residual deviance: 1.4655e-14  on 0  degrees of freedom
AIC: 25.476

Number of Fisher Scoring iterations: 3
```

(e) Figura 29. Resumen del modelo 4.

Finalmente vemos que para este modelo además del intercepto, ninguna de las variables resultó ser signifivativa en el momento del entrenamiento, lo cual podría indicar que la correlación entre las dos variables no es tan fuerte como se cree por lo cual una es la que mejor predice. Para este modelo el coeficiente de McFadden fue de 0.26.

Como conclusión se tiene que la mejor variable predictora para el modelo de regresión logística fue la AZT, lo cual puede indicar que no importa la raza de la persona, más bien si esta consume o no la medicina para la presencia de sida. El mejor modelo obtenido basado en el resumen de la fase de entrenamiento fue el que únicamente incluia como variable predictora la AZT.