k-Nearest Neighbors

Con objeto de ajustar el modelo a nuestra base de datos para predecir la variable respuesta, se usará el método kNN. Para poder ajustar el modelo de manera óptima se sigue un proceso de preparación de los datos, donde se dividen los en dos conjuntos: un conjunto de entrenamiento y un conjunto de prueba. El primer grupo que se utilizará para entrenar el modelo kNN estará compuesto por el 80 % de la base de datos original. Asimismo, el conjunto de prueba se empleará para evaluar el rendimiento y precisión del modelo.

A continuación, en el siguiente codigo aparece la generación del conjunto de train y test:

```
set.seed(12345)
Index <- createDataPartition(y, p = 0.8, list = F)
```

La selección de un valor de K se considera un paso crucial, ya que K es un hiperparámetro en kNN que representa el número de vecinos más cercanos a considerar. Se recomienda realizar pruebas con diferentes valores de K y utilizar la validación cruzada para determinar el valor óptimo. La validación cruzada se usará dentro del conjunto de datos de entrenamiento para encontrar el valor óptimo de k.

En este caso, la realización de la Cross-validación se realiza a partir de folds. Esto consiste en dividir la base de datos perteneciente al entrenamiento en un número determinado de subgrupos aleatorios y ejecutar el algoritmo considerando como test un fold distinto en cada una de las iteraciones. En cada una de las iteraciones se calcula la precisión del algoritmo, para posteriormente calcular la media de estas precisiones.

El número de vecinos que haya tenido una media de las precisiones mayor, será el escogido.

Para el proceso de Cross-Validación se han fijado unos valores de k del 1 al 20. En cuanto al número de folds, se ha considerado oportuno utilizar una cantidad de 10 folds, lo que supone ejecutar el kNN 10 veces para cada uno de los valores de k propuestos. Esto hace que durante el proceso de Cross-Validación el kNN sea ejecutado una totalidad de 200 veces.

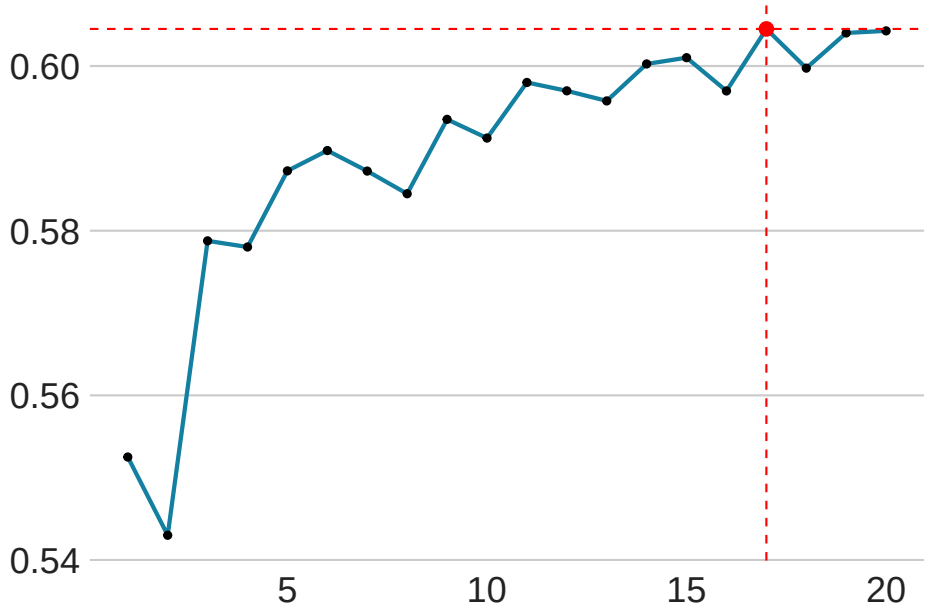
Acto seguido, se muestra una tabla en la que se recoge la media del accuracy para cada una de las k utilitzadas en el proceso de Cross-Validación.

Cuadro 50: Medias de Accuracy para cada k en la CV

	k	mean_acc
k= 1	1	0.5525113
k= 2	2	0.5430082
k= 3	3	0.5787646
k= 4	4	0.5780164
k= 5	5	0.5872664
k= 6	6	0.5897346
k= 7	7	0.5872390
k= 8	8	0.5844864
k= 9	9	0.5935195
k= 10	10	0.5912482
k= 11	11	0.5979995
k= 12	12	0.5969865
k= 13	13	0.5957607
k= 14	14	0.6002520
k= 15	15	0.6010109
k= 16	16	0.5969709
k= 17	17	0.6045015
k= 18	18	0.5997415
k= 19	19	0.6040159
k= 20	20	0.6042696

Con el fin de facilitar la interpretación se reproduce un gráfico de la tabla anterior. En este se resalta la k con la que se ha conseguido un Accuracy más elevado y seguidamente se muestra su valor.

Figura 124: Media del Accuracy para cada k



Como se puede ver, la k que ha conseguido una Accuracy más elevada es la siguiente:

Cuadro 51: Accuracy de la k optima en la CV

	k	mean_acc
k= 17	17	0.6045015

Una vez terminado el proceso de Cross-Validación y habiendo encontrado la k óptima, el siguiente paso implica la implementación del algoritmo (con la k seleccionada) para predecir la categoría de la variable respuesta de los individuos del test mediante la información que proporiconan los individuos del train.

Una vez ejecutado el kNN se muestra en una tabla la matriz de confusión y se calcula la precisión con la que el algoritmo ha predicho la variable Target en la población del test.

```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 418 231
         1 155 196

               Accuracy : 0.614
                 95% CI : (0.583, 0.6443)
    No Information Rate : 0.573
    P-Value [Acc > NIR] : 0.0046689

                  Kappa : 0.1929

 Mcnemar's Test P-Value : 0.0001349

            Sensitivity : 0.7295
            Specificity : 0.4590
         Pos Pred Value : 0.6441
         Neg Pred Value : 0.5584
             Prevalence : 0.5730
         Detection Rate : 0.4180
   Detection Prevalence : 0.6490
      Balanced Accuracy : 0.5943

       'Positive' Class : 0
```

La anterior salida nos muestra la matriz de confusión junto con diversos estadísticos que tratan de explicar como de bien o mal ha predicho el algoritmo de kNN.

De entre estos cabe destacar la Accuracy, que en este caso a sido de 0.614, por lo que el agloritmo ha predicho correctamente el 61.4 % de los individuos de Test.

La "Sensitivity" mide la proporción de individuos de TARGET=0 que han sido clasificacdos correctamente, que en este caso ha sido de NA.

Y finalmente la "Specificity" mide la proporción de individuos de TARGET=1 que han sido clasificados correctamente, que ha dado NA