|  |
| --- |
| **Machine Learning** |
| **Practice 2.6. Classification Hackathon** |
| **Authors:** |
| Data preprocessing:  Primero comprobamos si había valores ausentes, pero no había. Pasamos a buscar atípicos y encontramos un valor en la variable X8 bastante más grande de la media y lo eliminamos.  Además, miramos con ggpairs las correlaciones entre variables. Aquí nos dimos cuenta que X7 y X3 estaban muy correlacionadas (0.991). En un primer lugar, eliminamos la variable X7 del modelo.  Después, de está limpieza de datos empezamos con el análisis. Empezamos por modelos lineales y vimos que no nos daban buenos resultados. Añadimos alguna variable al cuadrado que parecía tener relación cuadrática en el ggpairs pero los resultados no fueron buenos. Pasamos al KNN y al Tree y mejoraba el rendimiento del modelo. El SMV\_Radial también nos daba buenos valores. Pero tras probar el MLP nos dimos cuenta de que sin duda es el mejor modelo para el análisis de estos datos.  Tras hacer varias pruebas cambiando parámetros y analizando la importancia de las variables del modelo nos dimos cuenta que el X7 era con diferencia la variable más importante. Por tanto, ya que tenia tanta correlación con X3 decidimos probar a realizar el análisis eliminando X7 en vez de X3.  Model comparison:   |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | | Model | Structure | Inputs | E training | E cross val | E validation | | MLP | size = 20 and decay = 0.01  iteraciones = 200 | x1, x2, x7, x9, x10 | Kappa : 0.98  Accuracy : 0.99 | Accuracy 0.9500266  Kappa  0.9000174 | Accuracy : 0.9548  Kappa : 0.9096 | | MLP | size = 5 and decay = 0.01  iteraciones = 200 | Todos -X3 | Accuracy : 0.9513  Kappa : 0.9026 | Accuracy  0.8950258  Kapa  0.7900434 | Accuracy : 0.8894  Kappa : 0.7789 | | SMV\_radial |  |  |  |  |  | |  |  |  |  |  |  | |  |  |  |  |  |  |   Conclusions:  Una vez decidimos que X7 era la variable más importante del conjunto de datos, ajustamos la red neuronal varias veces. Primero, eliminamos variables cuya importancia era muy limitada en los gráficos del modelo. En un primer lugar eliminamos X8 y X6. Los valores de precisión y de Kappa aumentaron así que decidimos eliminar también X3, X4 y X5 ya que también parecían poco relevantes. Con este conjunto de variables (x1, x2, x7, x9, x10) ajustamos el numero de neuronas y el parámetro decay, mediante el uso de cross validation y la ayuda del ggplot(mlp.fit)+scale\_x\_log10(). Finalmente, ajustamos el numero de iteraciones para evitar sobre-entrenar el modelo. Obtuvimos muy buenos resultados al comprobar nuestros datos en la pagina web. Un 94% con todas las variables menos la X3 y un 95% con x1, x2, x7, x9 y x10 |
|  |