

Clase 7: Proyecto DS

1. Importar el dataset 'housing.csv', 'persons.csv', 'persons_2.csv', 'persons_basquet.csv' y 'function.csv'
2. Obtener una visualización de los datos del dataset 'housing.csv'. (histograma, gráficos 2D/3D). Obtener valores máximos, mínimos, media, mediana, varianza, cross correlación.
3. Como parte de un pipeline, agregar 2 métodos distintos para completar los datos.
4. Agregar un transformador al pipeline que seleccione solo algunas features.
5. Dividir las muestras en datasets para entrenamiento del modelo, validación y test.
6. (Cuál es la diferencia entre el dataset de validación y el Dataset de test?)
7. Realizar una regresión polinomial para predecir el valor de $y=f(x)$ utilizando el dataset 'function.csv', variando el grado del polinomio.
 - a. Obtener los coeficientes del polinomio obtenido.
 - b. Obtener el error RMSE y MAE para el set de entrenamiento.
 - c. Obtener el error RMSE y MAE para el set de test.
8. Realizar un modelo que estime el sexo (0='M', 1='F') de una persona a partir de la altura y el peso. Utilizar el dataset 'persons.csv'.
 - a. Utilizar los datos del dataset 'persons_2.csv' para validar el modelo.
 - b. Considerar e implementar una estrategia para mejorar los resultados del predictor.
9. Crear un nuevo dataset uniendo los datasets: 'persons.csv', 'persons_2.csv' y 'persons_basquet.csv'.
 - a. Volver a entrenar el modelo y comparar la performance con el modelo del punto 8.
10. Realizar el entrenamiento del modelo utilizando Validación Cruzada (cross-validation)
11. A partir de los ejemplos visto en la demostración y utilizando el dataset 'housing.csv', probar entrenar un algoritmo SVM (sklearn.svm.SVR) con varios hyperparameters y comparar performance.
 - a. Probar con *kernel*=linear y varios valores de *C*.
 - b. Probar con *kernel*=rbf y varios valores de *C* y *gamma*.
12. A partir del punto anterior, aplicar *GridSearchCV* a etapas o pasos de la preparación de datos
13. Reemplazar *GridSearchCV* por *RandomizedSearchCV*.