Cuestionario general:

1. Establece con tus propias palabras, algunas buenas prácticas y funciones recomendadas para optimizar operaciones de lectura, escritura y manipulación en Spark/PySpark.
2. Indica las pruebas estadísticas que has utilizado como parte del desarrollo de una solución de ciencia de datos.

* Prueba de normalidad. Para verificar supuestos del modelo de regresión.
* Prueba de Durbin-Watson. Para probar independencia de los errores en el modelo lineal
* Prueba de Rachas. Para probar aleatoriedad en los errores.
* Prueba de White. Para probar la no existencia de heteroscedasticidad de los errores en el modelo lineal.
* Prueba de multicolinealidad (VIF) : También para validar supuesto del modelo línea, que no exista correlación entre las variables dependientes.
* Distancia de Cook. Para corroborar la no existencia de puntos de influencia que afecten los parámetros del modelo lineal.
* **Test de Grubbs. Para detectar la presencia de outliers en datos univariados**
* **Insolation Forest. Para detectar presencia de outliers en datos multivariados**
* **Prueba de Chi cuadrada. Para probar independencia entre variables categóricas.**

1. En el contexto de Machine Learning y Ciencia de datos, explica:
   1. No Free Lunch Theorem

En pocas palabras este teorema menciona que un algoritmo que funciona bien en cierto tipo de problema no tiene garantía de ser bueno en otro tipo diferente. Para unos datos puede funcionar mejor un random forest, mientras que para otro Redes neuronales, no existen un algoritmo que funcione bien para todos los datos (problemas) que existen.

Occam’s Razor.

Nos dice que se prefieren modelos simples que tienen el mismo rendimiento. Un modelo complejo, en general suele sobreajustar, lo cual genera bajos niveles de predicción cuando un modelo se pone en producción.

* 1. Data Leakage

Es un problema que surge cuando se incluye información que no estará disponible en el momento de hacer predicciones. Genera modelos que funcionan muy bien en entrenamiento, pero fallan en la práctica.