Examen diagnóstico

Alumno: Salazar Vega Rodrigo

1. ¿Qué es un banco de datos?
   Los datos que tenemos a disposición para trabajar.

2. En Python, ¿Qué es un dataframe?
   En python cuando nos referimos a un dataframe, nos referimos a una matriz de datos heterogeneos, o una matriz donde cada columna puede ser de un tipo de dato distinto a los demás.

3. ¿Qué son los valores perdidos?
   Son valores que o no se han registrado, que borraron o simplemente no lo tenemos en nuestro conjunto de datos.

4. ¿Qué se puede hacer si hay valores perdidos en tu banco de datos?
   Tenemos varias opciones, una seria que simplemente eliminaramos esos datos faltantes, esto seria viable si los datos faltantes representan un porcentaje muy pequeño de nuestros datos, si no, podríamos rellenar esos datos, ya sea con alguna métrica estadística, llamese moda, mediana, media o aplicando alguna técnica más compleja como la interpolacion o repetir los valos uno antes del dato faltante o uno después del dato faltante.

5. ¿Qué son los outliers?
   Son valores atípicos que son o muy grandes o muy pequeños respecto a nuestros datos o que se desvian más de tres desviaciones estandar de la media.

6. ¿Qué es media, mediana y moda?
   Son medidas estadísticas que representan varias cosas, la mediana es el valor que se encuentra justo en medio de nuestros datos ordenados, la media tambien es conocido como el valor esperado o el segundo cuartil, en otras palabras, es el promedio de nuestros datos y la moda es el valor que más se repite dentro de nuestros datos.

7. ¿Cuántos tipos de atributos pueden existir en un banco de datos?
   Tantos como el que lo diseño quiera

8. ¿Qué son las variables dummys?
   Son variables que sirven de auxiliar en algunas circustancias o que no afectan el aprendizaje de nuestro modelo.

9. ¿Cuál es la diferencia entre distribución uniforme y distribución normal?
   En una distribucion uniforme todos los números que tenemos ahí cuentan con la misma probabilidad de ocurrir, es decir, que ningun elemento de dicha distriubción es más probable que otro, mientras que la distribución normal no ocurre así, los valores más cercanos a la media o la media tienen una mayor probabilidad de ocurrir que los valores en los extremos, que por más pequeña probabilidad que tengan aún llegan a ocurrir, además la distribución normal se asemeja más a como funciona nuestro mundo

10. ¿Cuál es la diferencia entre aprendizaje supervisado, no supervisado y por reforzamiento?
    En el aprendizaje supervisado nosotros le damos al modelo una entrada y una salida que durante el entrenamiento el modelo puede o no acertar esa salida, mientras que en el no supervisado nostros solo le damos una entrada y el modelo nos brinda una salida que nostros no esperabamos o no sabiamos que podia salir y el aprendizaje por refuerzo el modelo mientras se esta entrenando al modelo se le dan diferentes estimulos para que cada iteración se comporte más como nostros queremos y no obtengamos una salida erronea.

11. ¿Cuándo se considera que existe un desbalance de clases?
    Cuando abunda más una clase sobre las demas, por ejemplo si tenemos 3 clases y tenemos las clases balanceadas, podríamos decir que cada una tiene $\frac{1}{3}$ del total de los datos, cuando no estan balanceadas una de ella podria tener $\frac{2}{3}$ o más y el resto de los datos estarían entre las dos clases restantes.

12. Menciona los métodos de validación que conoces para un banco de datos.
    Validación cruzada, k-pliegues.