Prueba de Conocimientos

Científico de Datos[1]

***Nota**: Lea atentamente toda la prueba antes de iniciar con su desarrollo. Si considera que falta algún dato o parámetro, asuma el valor que desee y justifíquelo.*

Problema 1 (1.5 Puntos)

Un 50 % de correos recibidos en un servidor llevan adjuntos y un 65 % son publicidad no deseada (SPAM). Sólo un 15 % de estos correos no llevan adjuntos y no son SPAM.

1. ¿Cuál es la probabilidad de que un correo lleve adjunto si es SPAM?
2. ¿Cuál es la probabilidad de que un correo no tenga adjuntos si no es SPAM?
3. Implemente una pieza de código[2] que reciba como input las probabilidades dadas en el enunciado y devuelva los valores pedidos en los literales anteriores.

Problema 2 (1.5 Punto)

La capacidad máxima de un ascensor es de $500kg$, por lo cual se recomienda que sea utilizado por un máximo de 6 personas simultáneamente. Los pesos de las personas que usan el ascensor tienen una distribución normal $\mathcal{N}(\mu, \sigma)$ con media μ y desviación estándar σ.

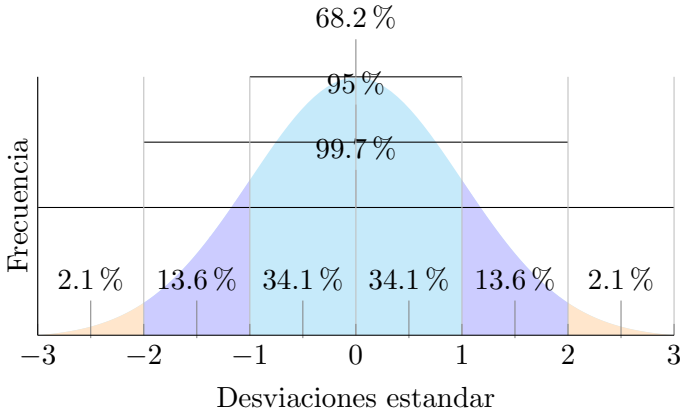

Figura 1: Distribución Normal

Para un valor de $\sigma = 15{,}7kg$ y de acuerdo a la figura 1.

(a) Cúal debe ser el valor de μ, para que la probabilidad de que las 6 personas en el ascensor superen su capacidad máxima sea del 2,1 %?

(b) Implemente una pieza de código para generar una lista aleatoria con distribución normal (μ, σ encontrados en el literal anterior) de los pesos en kg de las seis personas. Ejecútelo 10, 100, 1000 veces para encontrar experimentalmente en cada caso, la probabilidad de superar la capacidad del ascensor.

[1] Desarrollado por: Ángel Alberto Castro
[2] en el lenguaje de programación de su preferencia

Problema 3 (1 Punto)

Cómo científico de datos usted esta analizando un sistema de múltiples agentes heterogéneos, por ejemplo: clínicas, empresas de transporte (ambulancias), afectados (persona accidentada), proveedores de insumos médicos, vehículos con pólizas de seguro, siniestros, cuentas, etc. Usted nota que cada agente tiene ciertos atributos y que los agentes en general están fuertemente ligados por alguna relación. Por ejemplo, una persona accidentada estaba conduciendo un vehículo al momento del accidente, fue trasladado por una ambulancia a una clínica, en la clínica se le implantaron ciertos dispositivos médicos, etc. En su equipo de trabajo discuten una arquitectura para este sistema y se requiere a futuro poder consultar por ejemplo, cuales vehículos están asociados a un afectado?, A cuales clínicas ha llevado los pacientes una determinada ambulancia?, etc. De las siguientes opciones de AWS[3] cual considera más útil?. *justifique brevemente su respuesta*

(a) Amazon RDS
(b) Amazon RedShift
(c) Amazon Neptune
(d) Amazon ElasticSearch

Problema 4 (1 Punto)

Usted desea tener una base de datos que sea eficiente para realizar consultas analíticas sobre conjuntos grandes de datos columnares, y a la vez desea conectar sus consultas a un tablero de visualización para mostrar gráficas y reportes. Cual de las siguientes tecnologías de AWS considera más útil?. *justifique brevemente su respuesta*

(a) Amazon RDS
(b) Amazon S3
(c) Amazon Redshift
(d) Amazon Neptune
(e) AWS Elastic Beanstalk
(f) Ninguna de las anteriores

Problema 5 (1.5 Puntos)

Suponga que dispone de un archivo comprimido con información de eventos atmosféricos (tormentas) desde 1970 a 2022. Al cargar el archivo usted importa las siguientes columnas:

1. **episodioId**: Un único episodio puede contener múltiples eventos
2. **eventoId**: identificador único de cada evento
3. **departamento**: nombre de la región según mapa división política
4. **fechaInicial**: fecha y hora de cuando inicia el evento.
5. **torFscale** : escala Fujita-Pearson que mide la intensidad del evento (tornado) de acuerdo a la velocidad medida del viento. Inicia en F0 para eventos leves y va hasta F5 para eventos altamente destructivos.
6. **ubicacionOrigen**: nombre de la ciudad donde inicia el evento
7. **ubicacionDestino**: nombre de la ciudad donde termina el evento
8. **latitudInicial**: coordenada geográfica

[3] Amazon Web Services

9. **latitudFinal**: coordenada geográfica
10. **muertesTotales**: Cantidad de muertes atribuidas directa o indirectamente al evento.
11. **lesionesTotales**: Cantidad de personas lesionadas atribuibles al evento.
12. **impactoEstimado**: costo estimado en pesos de la reparación de los daños causados.

Indique qué tipo de visualizaciones utilizaría para explorar las distribuciones univariadas de cada una de las columnas y justifique brevemente los criterios para dicha selección. Puede sugerir más de una opción por variable o incluso simplemente justificar que no aplica visualización.

Problema 6 (2.5 Puntos)

En el sector asegurador, la capacidad para calcular las primas del seguro tiene un impacto significativo en la toma de decisiones de los directivos. Usted como científico de datos debe evaluar el estado actual para identificar factores relevantes. En el siguiente enlace encuentra un conjunto de datos: ***costos_primas.csv*** que ha sido preparado para los análisis descritos a continuación (la descripción de los campos se encuentra en la tabla 1):

https://testdatascientist-sis.s3.amazonaws.com/costos_primas.csv

Columna	Descripción
state	Estado (EUA) donde se compró la póliza
group_size	Número de beneficiarios
homeowner	Si el cliente posee o no casa propia
car_age	Tiempo en años del vehículo
risk_factor	Nivel de riesgo del usuario
age_oldest	Edad de la persona mayor en el grupo de beneficiarios
age_youngest	Edad de la persona más joven en el grupo de beneficiarios
married_couple	Hay pareja de casados en el grupo
C_previous	Tipo de póliza anterior (0:ninguna)
duration_previous	Tiempo en años que el usuario ha tenido pólizas
A,B,..,,G	Opciones de cobertura de la póliza
cost	Costo de la prima del seguro

Tabla 1: Descripcion atributos Problema 6

1. cargue el .csv en un dataframe de **pandas** y especifique las columnas A, B, C, D, E, F, G, car_value, state como variables categoricas.
2. Que otras variables categoricas identifica en el dataframe?
3. Convierta todas las variables categoricas para que queden en un formato de *one hot encoding*
4. Analice los valores extremos y faltantes para proponer un esquema de imputaciones y de limpieza adicional. Explique su propuesta y aplíquela al dataset.
5. Divida el conjunto de datos en un conjunto de entrenamiento y un conjunto de prueba (80-20). Use un valor de semilla de 1337
6. Ajuste un modelo de regresión lineal múltiple a los datos de entrenamiento siendo el costo la variable dependiente y todas las demás variables independientes.
7. Encuentre los valores: R^2, AIC

8. De acuerdo al resultado de su modelo de regresión cuales estados son los mas y los menos costosos (utilice un top 5)
9. Interprete los coeficientes de ***homeowner*** y ***car_age***. Los signos y valores de estos coeficientes tienen sentido para usted en el contexto del problema?
10. Cuales variables de su modelo son **estadísticamente significativas**?. Nota: Considere un nivel de significancia $\alpha = 0{,}01$. Para las variables categóricas, considérelas significativas si al menos una de sus categorías es significativa.

Problema 7 (1 Punto)

Usted tiene un dataset de aprox 6,000 registros con dos atributos: **edad_empleado** (rango de 20 a 60) y **salario_anual** (rango de $\$10'000{,}000$ a $\$180'000{,}000$). Cúal de las siguientes situaciones es **MÁS** probable de ocurrir si se ejecuta en *python* un proceso de clustering usando **k-means**?. *justifique brevemente su respuesta*

(a) Los datos seran agrupados apropiadamente
(b) El programa tendrá problemas de memoria RAM
(c) El programa tendrá problemas de desbordamiento numérico
(d) Los *clusters* no proporcionarán información útil
(e) Algunos *datapoints* pertenecerán a más de un cluster
(f) Ninguna de las anteriores

Entregable

- Si copia código de algún lugar, por favor referencielo.
- Comente su código fuente brevemente en donde considere necesario.
- Use GIT para el versionamiento del proyecto (enviar link del repositorio).
- En el repositorio: realice al menos dos commits
- Archivo de texto con sus respuestas y justificaciones (readme del repositorio, o pdf, o word)