Actividad Evaluable 1

ALUMNOS:

- DIEGO SOSA ALVA
- FELIPE VASQUEZ VASQUEZ

Data Science

1. Pregunta 1

1.1. Dado un registro de vehículos que circulan por una autopista, disponemos de su marca y modelo, país de matriculación, y tipo de vehículo (por número de ruedas). Con tal de ajustar precios de los peajes, ¿Cuántos vehículos tenemos por tipo? ¿Cuál es el tipo más frecuente? ¿De qué países tenemos más vehículos?

RPTA:

- **Descriptiva:** Porque el texto describe los datos disponibles en los registros de vehículos que circulan por la autopista.
- **Exploratoria:** Porque el propósito de una de las preguntas es explorar los datos disponibles para determinar la cantidad de vehículos por tipo.
- **Exploratoria:** Porque el propósito de una de las preguntas es explorar los datos disponibles para determinar qué vehículo es el más frecuente.

1.2. Dado un registro de visualizaciones de un servicio de video-on-demand, donde disponemos de los datos del usuario, de la película seleccionada, fecha de visualización y categoría de la película, queremos saber ¿Hay alguna preferencia en cuanto a género literario según los usuarios y su rango de edad?

RPTA:

- **Descriptiva:** Porque el texto que describe los datos del historial de visualización disponibles para los servicios de video a pedido.
- **Exploratoria:** Porque el propósito de una de las preguntas es explorar los datos disponibles para determinar si existe una preferencia por algún género literario en función de los usuarios y su grupo de edad.

1.3. Dado un registro de peticiones a un sitio web, vemos que las peticiones que provienen de una red de telefonía concreta acostumbran a ser incorrectas y provocarnos errores de servicio. ¿Podemos determinar si en el futuro, los próximos mensajes de esa red seguirán dando problemas? ¿Hemos notado el mismo efecto en otras redes de telefonía?

RPTA:

- **Descriptiva:** Porque el texto que describe los datos disponibles en los registros de solicitud del sitio y los patrones de error asociados con una red de telefonía en particular.

- **Predictiva:** Porque el objetivo de una de las preguntas es predecir si el próximo mensaje de la red de telefonía seguirá causando problemas.
- **Inferencial:** Porque el propósito de una de las preguntas es concluir si se ha observado el mismo efecto en otras redes de telefonía.

1.4. Dado los registros de usuarios de un servicio de compras por internet, los usuarios pueden agruparse por preferencias de productos comprados. Queremos saber si ¿Es posible que, dado un usuario al azar y según su historial, pueda ser directamente asignado a un o diversos grupos?

RPTA:

- **Descriptiva:** Porque el texto describe los datos disponibles sobre el registro de usuarios de los servicios de compra online y la información que agrupa a los usuarios según sus preferencias de compra de productos.
- **Predictiva:** Porque el objetivo del problema es predecir si un usuario aleatorio se puede asignar directamente a uno o más grupos en función de su historial de compras.

2. Pregunta 2

Sabemos que un usuario de nuestra red empresarial ha estado usando esta para fines no relacionados con el trabajo, como por ejemplo tener un servicio web no autorizado abierto a la red (otros usuarios tienen servicios web activados y autorizados). No queremos tener que rastrear los puertos de cada PC, y sabemos que la actividad puede haber cesado. Pero podemos acceder a los registros de conexiones TCP de cada máquina de cada trabajador (hacia donde abre conexión un PC concreto). Sabemos que nuestros clientes se conectan desde lugares remotos de forma legítima, como parte de nuestro negocio, y que un trabajador puede haber habilitado temporalmente servicios de prueba. Nuestro objetivo es reducir lo posible la lista de posibles culpables, con tal de explicarles que por favor no expongan nuestros sistemas sin permiso de los operadores o la dirección.

RPTA:

Para iniciar, debemos definir el problema, este surge porque se observa actividad no autorizada de un usuario en la red permitiendo generar posibles brechas de seguridad. Por tanto, el objetivo es localizar al usuario en cuestión y tomar medidas para evitar que esto vuelva a suceder.

En la recopilación de datos se tiene las conexiones TCP de las máquinas de los usuarios, donde figuran las conexiones externas.

Luego pasamos a la limpieza y transformación de estos datos (preparar los datos), donde se podría optar por eliminar los registros considerados como legítimos y quedarse con los datos que puedan tener relación con el caso. Esto con la finalidad de reducir la información y hacerla más manejable.

Entonces pasamos a plantearnos las preguntas que podrían ser: ¿Quién realizó conexiones a lugares no autorizados durante el período en el que se detectó la actividad no autorizada? ¿Qué patrones se pueden identificar en los registros de conexiones? ¿Existen conexiones a los mismos lugares de forma recurrente? ¿Existen conexiones a sitios web no seguros o potencialmente

maliciosos? También se podrían plantear hipótesis sobre quién podría ser el posible culpable, basadas en la ubicación del PC, su rol en la empresa, o su historial de conexiones previas.

Ahora, con los datos preparados y las preguntas planteadas, pasamos al análisis donde utilizamos técnicas exploratorias y minería de datos. El objetivo es identificar patrones y anomalías para generar un modelo predictivo para ver la probabilidad de que una máquina este siendo utilizada para actividad no autorizada.

Por último, debemos presentas estos resultados a los interesados utilizando gráficos de perfil ejecutivo y entendido a todo nivel. Aquí se debe exponer las medidas preventivas para evitar casos similares en el futuro.