Diplomado en ciencia de datos
Hackaton de Analítica
Proyecto Final

Objetivo: Resolver la mayor cantidad de retos propuestos en las diferentes bases de datos.

Reglas:

- Cada equipo se conforma por dos personas
- Cada reto tiene una puntuación máxima de acuerdo con su dificultad, la puntuación final por equipo será la suma de los puntos obtenidos por cada reto
- Por cada reto resuelto se debe tener:
    - Una o dos diapositivas explicando cómo resolvió el reto y la conclusión a la que llegó (si el reto es muy corto, inclúyalo con varios retos cortos en una sola diapositiva).
    - El código que fue empleado ya sea en python o en R
- Los equipos pueden obtener bonificaciones si utilizan las bases de datos para resolver preguntas que no hagan parte de los retos
- En cada reto el equipo debe decidir entre una de las siguientes categorías para el reto:

    A - Completado al 100%
    B - Completado en un 75%
    C - Completado a la mitad
    D - Completado en un 25%
    E - Reto no iniciado

Retos:

Se cuenta con un conjunto de bases de datos con diferentes objetivos de análisis y diferentes tipos de datos. Por cada una de las bases de datos son propuestos varios retos que pondrán a prueba las habilidades de:

- Limpiar y pre-procesar la información antes de aplicar determinada técnica
- Identificar el tipo de análisis pertinente de acuerdo con la información disponible y el objetivo de este
- Procesar la información de forma adecuada
- Interpretar los resultados de los análisis
- Capacidad de transmitir los resultados del análisis de la forma correcta

Primera fuente de datos

Factores de riesgo de cáncer cervical

El cáncer cervical es el cáncer que comienza en el cuello uterino, la parte inferior del útero (matriz) que desemboca en la parte superior de la vagina. El desarrollo del cáncer cervical generalmente es muy lento y comienza como una afección precancerosa llamada displasia. Esta afección se puede detectar por medio de una citología vaginal y es 100% curable. Pueden pasar años para que la displasia se convierta en cáncer cervical. La mayoría de las mujeres a quienes se les diagnostica cáncer cervical en la actualidad no se han sometido a citologías vaginales regulares o no han tenido un seguimiento por resultados anormales en estas. Casi todos los cánceres cervicales son causados por el virus del papiloma humano (VPH).

Se desea establecer una relación entre el diagnostico de cáncer de cuello uterino y diferentes variables que pueden afectar la aparición de este. En el siguiente link se puede acceder a una base de datos abiertos recolectados en el Hospital Universitario

de Caracas, Venezuela. El conjunto de datos incluye información demográfica, hábitos y registros médicos históricos de 858 pacientes.

Observación: El conjunto de datos tiene datos perdidos, ya que un porcentaje de la población decidió no responder a algunas preguntas por problemas de privacidad.

Variables disponibles:

Variable	Tipo de dato	Nombre en la base de datos
Edad	Entero	edad
Número de parejas sexuales	Entero	parejas_sexuales
Edad de la primera relación sexual	Entero	primera_relacion
Número de embarazos	Entero	numero_embarazos
Fuma (1 = si, 0 = no)	Binaria	fuma
Número de años fumando	Entero	anios_fumando
Numero de paquetes por año	Entero	paquetes_anio
Uso de anticonceptivos (1 = si, 0 = no)	Binaria	anticonceptivos
Número de años usando anticonceptivos	Entero	anios_anticonceptivo
Uso de dispositivo intrauterino - DIU (1 = si, 0 = no)	Binaria	diu
Número de años usando DIU	Entero	diu_anos
Enfermedad de transmisión sexual (1 = si, 0 = no)	Binaria	ets
Número de enfermedades de transmisión sexual	Entero	ets_numero
Condilomatosis (1 = si, 0 = no)	Binaria	ets_condilomatosis
Condilomatosis cervical (1 = si, 0 = no)	Binaria	ets_condilomatosis_cervical
Condilomatosis vaginal (1 = si, 0 = no)	Binaria	ets_condilomatosis_vaginal
Condilomatosis vulvo perineal (1 = si, 0 = no)	Binaria	ets_condilomatosis_vulvo_perineal
Sífilis (1 = si, 0 = no)	Binaria	ets_sifilis
Enfermedad inflamatoria pélvica (1 = si, 0 = no)	Binaria	ets_enfermedad_inflamatoria_pelvica
Herpes genital (1 = si, 0 = no)	Binaria	ets_herpes_genital
Molluscum contagiosum (1 = si, 0 = no)	Binaria	ets_molluscum_contagiosum
SIDA (1 = si, 0 = no)	Binaria	ets_sida
VIH (1 = si, 0 = no)	Binaria	ets_vih
Hepatitis B (1 = si, 0 = no)	Binaria	ets_hepatitis_b
Virus del papiloma humano (1 = si, 0 = no)	Binaria	ets_vph
Número de diagnóstico	Categórica	numero_diagnostico
Años transcurridos desde el primer diagnóstico	Entero	tiempo_desde_primer_diagnostico
Años transcurridos desde el último diagnóstico	Entero	tiempo_desde_ultimo_diagnostico
Diagnóstico de cáncer (1 = si, 0 = no)	Binaria	diag_cancer
Diagnóstico de neoplasia cervical intraepitelial (1 = si, 0 = no)	Binaria	diag_neoplasia_cervical_intraepitelial
Diagnóstico virus de papiloma humano (1 = si, 0 = no)	Binaria	vph
Diagnóstico	Binaria	diag
Prueba hinselmann	Binaria	hinselmann
Prueba Schiller	Binaria	schiller
Prueba citología	Binaria	citologia
Prueba biopsia	Binaria	biopsia

Práctica

[Reto 1 – Limpiando y organizando la base de datos – 10pts]

Caracterizar los datos faltantes

- ¿Qué tipo de información es considerada por los pacientes cómo información sensible?
- ¿Existen variables con una cantidad grande de datos faltantes?
- ¿Qué porcentaje de los datos son faltantes en total y por variable?

[Reto 2 – Resumiendo las variables – 30pts]

- ¿Cuál es el número esperado de parejas sexuales de una mujer y desde que cantidad podría decirse que es "atípico"?
- ¿Existen datos que puedan considerarse atípicos? ¿qué característica tienen dichos datos?
- ¿Existen diferencias estadísticamente significativas en las edades de las mujeres que fueron diagnosticadas con cáncer frente a las que no? Concluya con un 95% de confianza
- ¿Qué puede concluir acerca de una posible relación entre ser fumador y tener cáncer?

[Reto 3 – Modelando la aparición del cáncer – 60pts]

Haga un modelo adecuado que permita predecir, con base en los datos cuales son las características que más influyen en la aparición de cáncer cervical

[Reto 4 – Construyendo una variable latente – 70pts]

Genere un modelo que le permita cuantificar la salud sexual de una paciente

Segunda fuente de datos

Medidas de interacción de usuarios en Facebook

La base de datos contiene la información relacionada con 500 post de Facebook hechos en el 2014 de la página de una empresa de productos cosméticos. link

Variable	Tipo de dato	Nombre en la base de datos
Total «me gusta» de la página (Recordar que es el número de «me gusta» de la página de cosméticos)	Entero	total_likes_pagina
tipo (foto, link, video, status)	Entero	tipo
Categoría (entre: 1 , 2 , 3)	Categórica	categoria
Mes en que fue realizado el post	Ordinal	mes
Dia de la semana en que fue realizado el post	Ordinal	dia
Hora en que se realizó el post	Hora	hora
Pago	Binaria	pago
Número de personas que vieron una publicación de la página (usuarios únicos)	Entero	personas_vieron
Número de veces que se muestra una publicación de una página - (Impresiones)	Entero	numero_impresiones
Número de usuarios comprometidos (usuarios únicos que le dieron click en algún lugar de la publicación)	Entero	usuarios_comprometidos_unicos
Número de usuarios comprometidos (no únicos)	Entero	usuarios_comprometidos
Numero de clicks en cualquier lugar de una publicación	Entero	numero_cliks
Numero de impresiones a usuarios que le dieron «me gusta» a la página	Entero	usuarios _like_ impresiones
Numero de usuarios únicos que han visto la publicación por que le dieron «me gusta» a la pagina	Entero	usuarios_like_vieron
Numero de usuarios únicos que se comprometen con la publicación porque le dieron «me gusta» a la pagina	Entero	usuarios_like_comprometidos
Cantidad de comentarios sobre la publicación	Entero	comentarios
Cantidad de «me gusta» en una página	Entero	likes
Cantidad de veces que se ha sido compartida la publicación	Entero	compartida
Suma de «me gusta», comentarios y acciones de la publicación	Entero	interacciones

Práctica

[Reto 1 – Limpiando y organizando la base de datos – 10pts]

Caracterizar los datos faltantes

- Genere un indicador único para cada una de las publicaciones de la base de datos
- ¿Cuántos datos faltantes hay y de que variables son?
- Para efectos del análisis borre los datos faltantes

[Reto 2 – Resumiendo las variables – 50pts]

- Haga un gráfico que le permita ver cómo ha sido el comportamiento de la cantidad de <<me gusta>> a lo largo del tiempo. Concluya acerca de la tendencia
- ¿Cuál es el número promedio de interacciones, impresiones y usuarios comprometidos dependiendo del tipo de publicación?
- Caracterice el número de veces en promedio un usuario que le ha dado «me gusta» a la página se comprometen con una publicación frente al total de usuarios comprometidos. Genere más métricas que le permiten describir los usuarios que le han dado <<me gusta>> a la página

[Reto 3 – Modelos – 70pts]

La variable “categoría” es una categoría creada por uno de los analistas, ¿puede reconstruir un posible significado de dicha categoría basándose en los datos disponibles? ¿Qué puede concluir de la relación entre el número de interacciones y las demás características de la publicación?

[Reto 4 – Agrupando – 60pts]

Genere una posible agrupación de las publicaciones de Facebook ¿Qué sugieren estos grupos?

Tercera fuente de datos

Desempeño de los estudiantes

Este es un conjunto de datos educativos que se recopilan a partir del sistema de gestión de aprendizaje (LMS) llamado Kalboard 360. Kalboard 360 es un LMS multiagente, que se ha diseñado para facilitar el aprendizaje mediante el uso de tecnología de vanguardia. Dicho sistema proporciona a los usuarios un acceso sincrónico a los recursos educativos desde cualquier dispositivo con conexión a Internet.

El conjunto de datos consta de 480 registros de estudiantes y 16 características. Las características se clasifican en tres categorías principales: (1) características demográficas como el género y la nacionalidad. (2) Características académicas básicas como la etapa educativa, el nivel de grado y la sección. (3) Características de comportamiento tales como la mano levantada en la clase, los recursos de apertura, la encuesta de respuesta de los padres y la satisfacción escolar.

Variable	Tipo de dato	Nombre en la base de datos
Genero	Nominal	genero
Nacionalidad	Nominal	nacionalidad
Lugar de Nacimiento	Nominal	lugar_nacimiento
Nivel Educativo	Nominal	nivel_educativo
Grado actual	Nominal	grado
Sección ID	Nominal	seccion
Tema	Nominal	tema
Semestre	Nominal	semestre
Familiar responsable	Nominal	familiar_responsable
Número de veces que el estudiante levanta la mano	Entero	participacion_mano
Número de veces que el estudiante usa los recursos	Entero	participacion_recursos
Número de veces que el estudiante ve los anuncios	Entero	participacion_anuncios
Número de veces que el estudiante participa en los foros	Entero	participacion_grupos
El familiar responde la encuesta	Binaria	encuesta_familiar
Satisfacción del pariente	Binaria	satisfaccion_familiar
Número de días que se ausenta de clases (mas o menos de 7 días)	nominal	dias_ausencia

Práctica

[Reto 1 – Limpiando y organizando la base de datos – 10pts]

- Caracterizar los datos faltantes
- ¿Cuántos datos faltantes hay y de que variables son?
- Para efectos del análisis borre los datos faltantes

[Reto 2 – Resumiendo las variables – 30pts]

- Existen medidas descriptivas que permitan observar si existen condiciones sociodemográficas que afecten la forma en la que el estudiante interactúa con la aplicación.
- Existen estudiantes que puedan ser considerados "atípicos" ¿qué características tienen estos estudiantes?

[Reto 3 – Modelos – 60pts]

Establezca una relación entre la satisfacción de los padres y las condiciones sociodemográficas y la interacción de los estudiantes con la aplicación. ¿Qué conclusiones tiene al respecto?

[Reto 4 – Agrupando]

Hay evidencia de la existencia de "grupos" de estudiantes que puedan ser atribuidos a las características observadas. ¿Cómo se caracterizan dichos grupos?

Cuarta fuente de datos

Prevalencia del consumo de sustancias psicoactivas

El siguiente link, contiene información de una base de datos abiertos de Colombia donde se presentan las cifras dadas por el ministerio de justicia acerca de la prevalencia del consumo de sustancias psicoactivas en la población nacional de Tabaco, Alcohol, Marihuana, Cocaína, Heroína, Solventes y Cualquier sustancia ilícita según dominio departamental.

Práctica

[Reto único – 120pts]

- Genere grupos de departamentos de acuerdo con el consumo de sustancias psicoactivas
- Utilice la siguiente base de datos (link) para cruzar la información del hurto a personas con respecto al consumo de sustancias psicoactivas ¿Es posible que exista una relación entre estos dos factores?

Quinta fuente de datos

Índice de felicidad por países

El Informe de la felicidad mundial es una encuesta histórica del estado de la felicidad global. Los informes revisan el estado de felicidad en el mundo actual y muestran cómo la nueva ciencia de la felicidad explica las variaciones personales y nacionales en la felicidad. Reflejan una nueva demanda mundial de más atención a la felicidad como un criterio para la política gubernamental.

Resultado del informe de la felicidad mundial, se tiene el ranking por países y la contribución que cada una de las variables que afectan un país tienen en la construcción de dicho índice. Los datos para 2015, 2016 y 2017 están disponibles en el siguiente link.

Práctica

[Reto 1 – Organizando la base de datos – 30pts]

- Cruce las bases de datos de 2015 y 2017 por cada uno de los países
- ¿Qué países fueron incluidos en los dos años? ¿Qué países fueron incluidos en solo uno de los dos años?
- Que puede concluir de la concordancia en el ranking de los países ¿Existen países con un cambio significativo en el ranking?

[Reto 2 – Analizando el índice construido – 100pts]

Concluya acerca de la consistencia del índice construido comparando tanto la contribución de las variables en la construcción del índice de felicidad como en la estabilidad de dichas contribuciones a lo largo de los dos años.

Cruce la información del índice de felicidad con la base de datos de las naciones unidas "UN" de la librería "carData" y genere una agrupación de países y un índice al que pueda darle interpretación

[Reto 2 – Analizando el índice construido – 80pts]

Haga un modelo que ayude a predecir la felicidad de las personas respecto a las demás variables de la base de datos de la ONU ¿Qué puede concluir al respecto?