Examen Final

Lea con atención las instrucciones que se le presentan. De manera individual, desarrolle las siguientes actividades. Tiene 90 min para responder, utilice su criterio y conocimientos desarrollados durante el curso para ejecutar cada actividad. Al finalizar el examen, deberá enviar por la plataforma Moodle un zip que contenga los elementos siguientes:

1. Word con respuestas de Serie 1
2. PDF serie 2
3. Script de R con el código generado (no olvide comentar su procedimiento)
4. Archivo de Tableau (.twbx)

**Serie 1 (20 puntos)**

1. Menciones por lo menos, 4 algoritmos de clasificación
2. Menciones por lo menos 3 metodologías de ciclo de vida de datos vistas en clase
3. ¿Cuál es la diferencia entre el soporte y la confianza? y en qué tipos de algoritmos de machine learning se utiliza.
4. ¿Qué es inteligencia de negocios?
5. Defina con sus propias palabras, un problema de inteligencia de negocios y cómo lo resolvería con las técnicas aprendidas en clase.

**Serie 2 (60 puntos)**

Para este ejercicio, debe de utilizar el dataset final.csv. El set de datos incluye datos demográficos de los 50 estados de Estados Unidos, con las siguientes columnas:

* Population: population estimate as of July 1, 1975.
* Income: per capita income (1974)
* Illiteracy: illiteracy (1970, percent of population)
* Life Exp: life expectancy in years (1969-71)
* Murder: murder and non-negligent manslaughter rate per 100,000 population (1976)
* HS Grad: percent high-school graduates (1970)
* Frost: mean number of days with minimum temperature below freezing (1931-1960) in capital or large city
* Area: land area in square miles

El objetivo de este ejercicio es que usted puede estimar la variable esperanza de vida (life expectancy – life Exp) a partir de las otras variables del set de datos (si aplicaran). Para ello debe de aplicar los conceptos vistos en clase:

* Correlación entre variables
* Validación cruzada (cross validation)
* Feature engineering
* Regresiones lineales

Debe crear un modelo de regresión lineal que nos brinde el menor RMSE de la combinación de variables que usted en base a su percepción y al análisis numérico del set de datos determine pertinentes. Se evaluará el trabajo de ingeniería efectuado (recuerde somos ingenieros) con los datos y como determino usted el mejor modelo. Por temas de tiempo, le recomiendo efectuar un análisis de al menos tres modelos diferentes de regresión y utilizar el que crea usted más conveniente.

Entregables:

* PDF detallando las decisiones que tomo al momento de efectuar el proceso de ingeniería de características, el cross validation y las regresiones lineales.
* Código de su solución.

**Serie 3 (20 puntos)**

Utilice los documentos CSV contenidos en la carpeta “Dashboard” para armar un modelo que permita la extracción de información. Debe verificar los datos contenidos, limpiarlos y seleccionar los requeridos para construir un dashboard que responda a las preguntas de negocio siguiente siguiente:

1. ¿Cuáles son las ventas totales por producto?
2. ¿Cuál es la venta total de la compañía en el año 2014?
3. ¿Qué talla y color de producto es el más vendido?