José Carlos Girón Márquez

1064718

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. Dashboard PowerBI
3. Script de R con el código generado (no olvide comentar su procedimiento)
4. Presentación de PowerPoint

**Serie 1 (10 puntos)**

1. **Menciones por lo menos, 4 algoritmos de clasificación**

* Regresión Logística
* Árbol de decisión
* K Nearest Neighbors
* Random Forest

1. **Menciones por lo menos 3 metodologías de ciclo de vida de datos vistas en clase**

* CRISP-DM
* KDD (Knowledge Discovery in Databases)
* SEMMA (Sample, Explore, Modify, Model, Assesment)

1. **¿Cuál es la diferencia entre el soporte y la confianza? y en qué tipos de algoritmos de machine learning se utiliza**

El soporte indica la cantidad de subconjuntos (itemset) se encuentra entre todas las transacciones, es decir, una probabilidad absoluta. Mientras que la confianza mide la frecuencia de aparición de los ítems en Y en transacciones que contienen X, en otras palabras, es una probabilidad condicional.

Puede encontrarse en algoritmos no supervisados como en las reglas de asociación.

1. **¿Qué es inteligencia de negocios?**

Business Intelligence es el proceso de convertir datos en conocimiento y conocimiento en acción para la toma de decisiones. Posee la característica de que la información es accesible, ayuda en la toma de decisiones y está orientado al usuario final. Utiliza herramientas tales como cubos OLAP, visualizadores o incluso consultas simples (queries).

1. **Defina con sus propias palabras, un problema de inteligencia de negocios y cómo lo resolvería con las técnicas aprendidas en clase.**

Se me viene a la mente el Marketing, el cual es un área que requiere de una gran cantidad de análisis de datos para buscar la mejor manera de dar a conocer sus productos. En primer lugar, se debe comprender el negocio ¿Qué es lo que busca vender? Luego, se debe de comprender qué datos se tienen disponibles, por ejemplo, informes de ventas. Posteriormente, se deben preparar los datos para ser utilizables. Luego, viene la parte del modelado en la que se puede usar uno de los numerosos algoritmos vistos, por ejemplo, el market basket análisis haciendo uso de reglas de asociación. Después del modelado se debe evaluar qué tan efectivo es y repetir hasta que se esté satisfecho, por último, implementar, tomar las predicciones y lanzar el producto en base a lo aprendido.

**Serie 2 (15 puntos)**

Utilice el documento examen.pbix para desarrollar los elementos siguientes:

1. Complete el modelo de extracción agregando la dimensión de tiempo

Graphical user interface, application

Description automatically generated

1. Elabore las siguientes columnas calculadas
   1. Año
   2. Mes

Graphical user interface, table

Description automatically generated

1. Elabore las siguientes métricas
   1. Ticket promedio (ventas / transacciones)

Graphical user interface, text, application

Description automatically generated

* 1. Ventas totales

Graphical user interface, application

Description automatically generated

* 1. YoY (cambio porcentual de año vrs año anterior)

Graphical user interface, application

Description automatically generated

1. Cree un dashboard utilizando los elementos siguientes
   1. Buenas prácticas de visualización
   2. Gráfico de líneas
   3. Gráfico de tarjeta
   4. Gráfico de pie

Chart, bar chart

Description automatically generated

**Serie 3 (20 puntos)**

En la siguiente serie deberá utilizar las fuentes de datos indicadas para analizar la información usando R.

La empresa “Bancopoli” es un respetado banco guatemalteco. Poseen una base de datos que contiene la información de sus clientes principales y su uso de productos financieros por medio del canal digital. Se le proporciona un pequeño dataset llamado “BankOnline.csv” que contiene una muestra de los datos con la información siguiente:

* Id-identificador de cliente
* Profit- ganancias por cliente
* Inc-Ingreso de los clientes
* Marge-margen de ganancia
* District-distrito donde vive
* Online-boolena, es cliente online o no

Por lo que le solicita los siguiente:

* 1. Genere la estadística General de los datos.

A screenshot of a computer

Description automatically generated with medium confidence

* 1. Debe de realizar un script que permita ejecutar los algoritmos siguientes para determinar si un cliente utilizará o no el canal Online:
     1. Decision Trees (aplicar poda)
     2. Random Forest
     3. Support Vector Machine
  2. Realice la evaluación de cada uno de los modelos mediante los elementos:
     1. Realice la matriz de confusión
     2. Realice el gráfico ROC

**Árboles de decisión**

Diagram

Description automatically generated

A picture containing text

Description automatically generated

Chart, line chart

Description automatically generated

**Random Forest**

A picture containing graphical user interface

Description automatically generated

Chart

Description automatically generated

**SVM**

A picture containing table

Description automatically generated

Chart, scatter chart

Description automatically generated

Chart, scatter chart

Description automatically generated

**Graphical user interface, chart

Description automatically generated with medium confidence**

**Serie 4 (35 puntos)**

En una presentación de PowerPoint, basada en sus habilidades como analista de datos, desarrolle los siguientes elementos:

* 1. Análisis de los 3 algoritmos las matrices de confusión
  2. Coloque las distintas gráficas ROC
  3. Establezca que algoritmo elegirá para desarrollar su proyecto y el motivo de seleccionarlo.
  4. Explique que algoritmos de machine learning podría utilizar para predecir el profit de los clientes online y cómo lo implementaría.

\*\*\*Ver documento adjunto

**Serie 5 (20 puntos)**

Haciendo uso del dataset “BankOnline.csv” realice una reducción de las variables haciendo uso del algoritmo para análisis de componentes principales y determine lo siguiente:

* **Cuantos componentes principales utilizará para su análisis**

Se utilizarán 2 componentes: PC1 y PC2.

* **Qué porción acumulada de datos cubren sus componentes principales seleccionados**

0.47 y 0.8 Respectivamente

Text

Description automatically generated

Chart, line chart

Description automatically generated

Chart, scatter chart

Description automatically generated