Universidad Rafael Landívar
Facultad de Ingeniería
Análisis de Datos
Msc. Ing. Pablo Orellana

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

2. **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)

3. **¿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.

Universidad Rafael Landívar
Facultad de Ingeniería
Análisis de Datos
Msc. Ing. Pablo Orellana

4. **¿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).

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

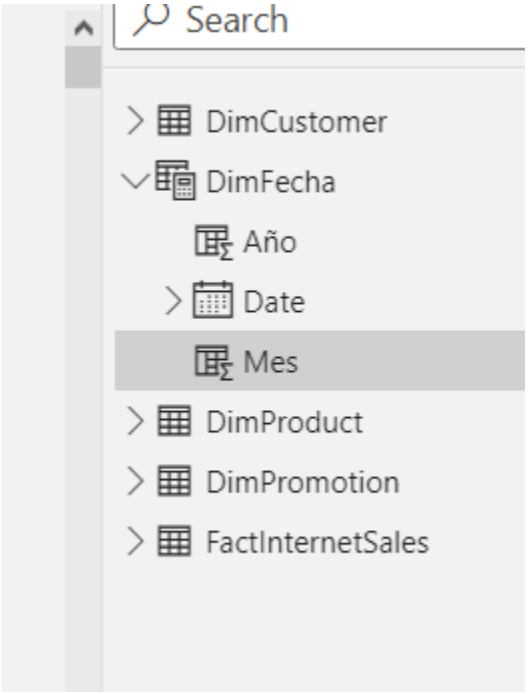
   Se me viene a la mente el Marketing, el cual es un área que requiere de una gran cantidad de análisis de datos para buscar la mejor manera de dar a conocer sus productos. En primer lugar, se debe comprender el negocio ¿Qué es lo que busca vender? Luego, se debe de comprender qué datos se tienen disponibles, 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:

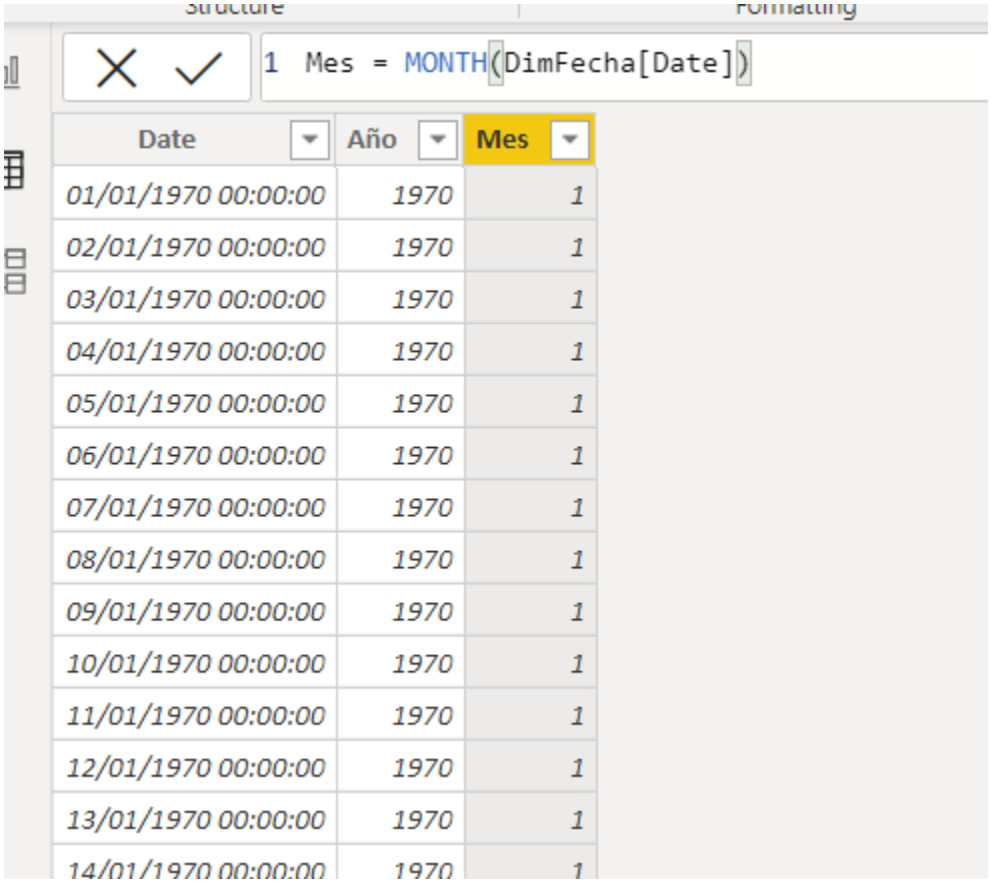
1. Complete el modelo de extracción agregando la dimensión de tiempo

Universidad Rafael Landívar
Facultad de Ingeniería
Análisis de Datos
Msc. Ing. Pablo Orellana



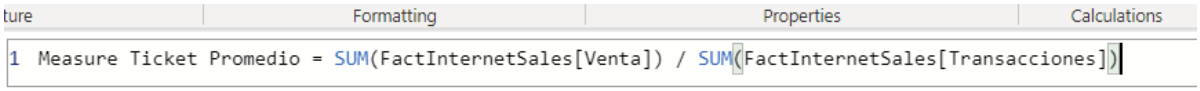
2. Elabore las siguientes columnas calculadas
    a. Año
    b. Mes

Universidad Rafael Landívar
Facultad de Ingeniería
Análisis de Datos
Msc. Ing. Pablo Orellana

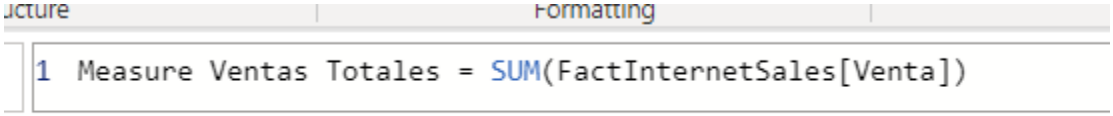


Date	Año	Mes
01/01/1970 00:00:00	1970	1
02/01/1970 00:00:00	1970	1
03/01/1970 00:00:00	1970	1
04/01/1970 00:00:00	1970	1
05/01/1970 00:00:00	1970	1
06/01/1970 00:00:00	1970	1
07/01/1970 00:00:00	1970	1
08/01/1970 00:00:00	1970	1
09/01/1970 00:00:00	1970	1
10/01/1970 00:00:00	1970	1
11/01/1970 00:00:00	1970	1
12/01/1970 00:00:00	1970	1
13/01/1970 00:00:00	1970	1
14/01/1970 00:00:00	1970	1

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



    b. Ventas totales



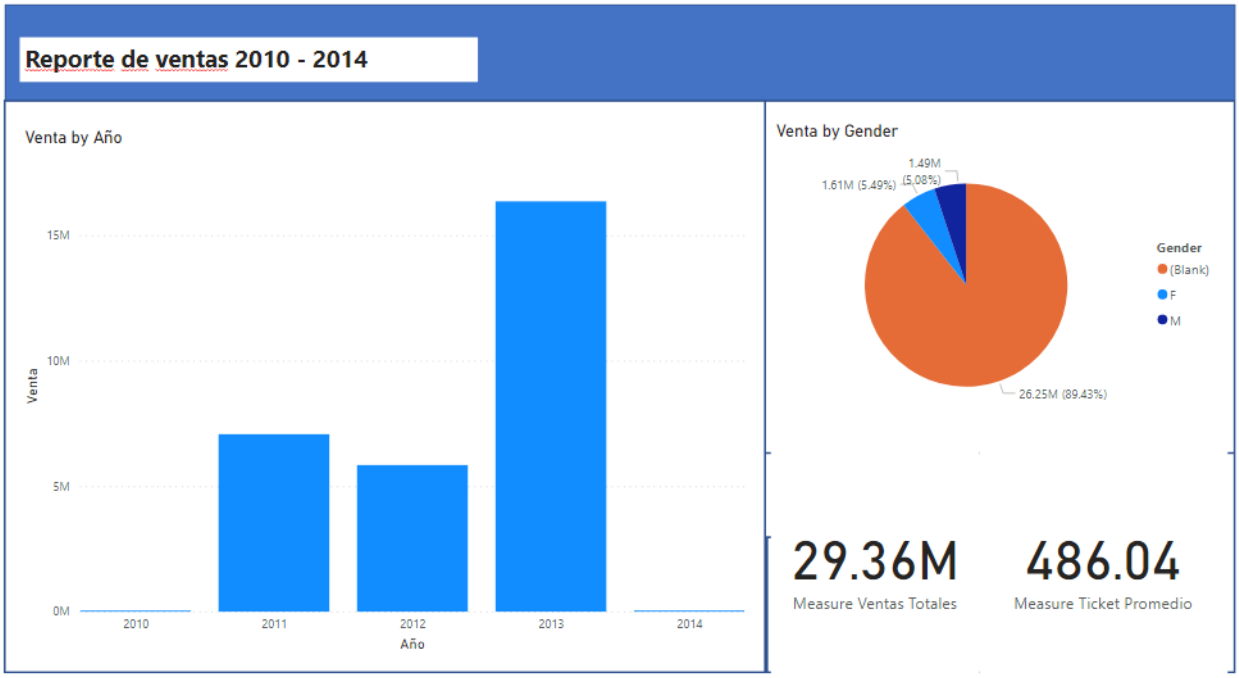
    c. YoY (cambio porcentual de año vrs año anterior)

Universidad Rafael Landívar
Facultad de Ingeniería
Análisis de Datos
Msc. Ing. Pablo Orellana

cture | Formatting

```
1  YoY = DIVIDE([ThisYear], [LastYear], 0)-1
```

4. Cree un dashboard utilizando los elementos siguientes
    a. Buenas prácticas de visualización
    b. Gráfico de líneas
    c. Gráfico de tarjeta
    d. Gráfico de pie



Universidad Rafael Landívar
Facultad de Ingeniería
Análisis de Datos
Msc. Ing. Pablo Orellana

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:

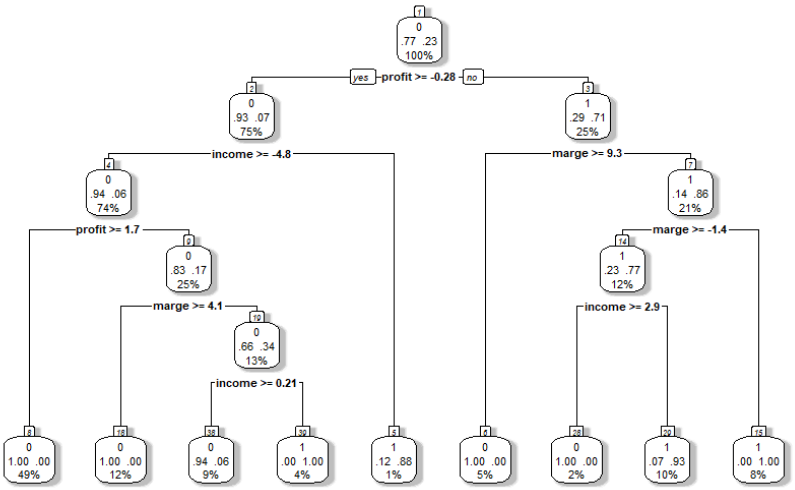
a. Genere la estadística General de los datos.

```
> #Estadística general de los datos
> summary(datos)
     profit             marge              income            entropy            online      
 Min.   :-6.5084   Min.   :-13.4979   Min.   :-5.2861   Min.   :-8.5482   Min.   :0.000  
 1st Qu.:-0.4566   1st Qu.: -0.3875   1st Qu.:-1.6662   1st Qu.:-2.2822   1st Qu.:0.000  
 Median : 1.5980   Median :  3.1650   Median : 0.6385   Median :-0.5634   Median :0.000  
 Mean   : 1.3339   Mean   :  3.0239   Mean   : 1.0795   Mean   :-1.1625   Mean   :0.238  
 3rd Qu.: 3.5265   3rd Qu.:  7.8663   3rd Qu.: 2.8646   3rd Qu.: 0.4090   3rd Qu.:0.000  
 Max.   : 6.8248   Max.   : 12.9516   Max.   :17.6772   Max.   : 2.4495   Max.   :1.000  
> 
```

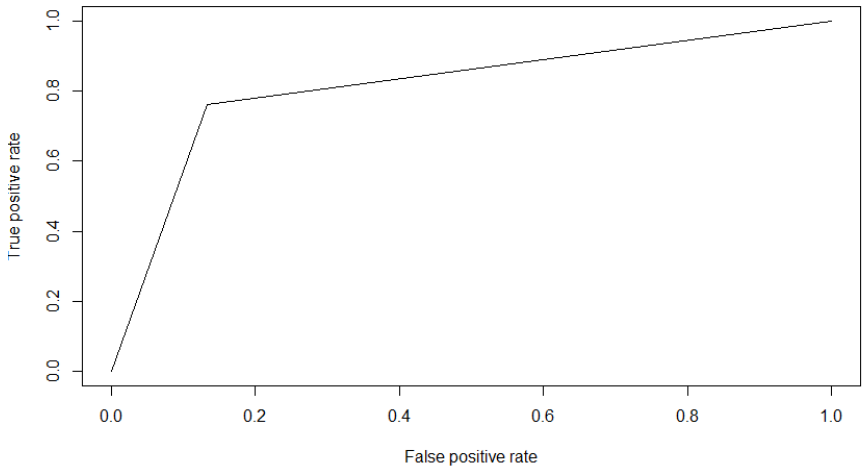
b. Debe de realizar un script que permita ejecutar los algoritmos siguientes para determinar si un cliente utilizará o no el canal Online:
    i. Decision Trees (aplicar poda)
    ii. Random Forest
    iii. Support Vector Machine
c. Realice la evaluación de cada uno de los modelos mediante los elementos: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

Universidad Rafael Landívar
Facultad de Ingeniería
Análisis de Datos
Msc. Ing. Pablo Orellana

Árboles de decisión



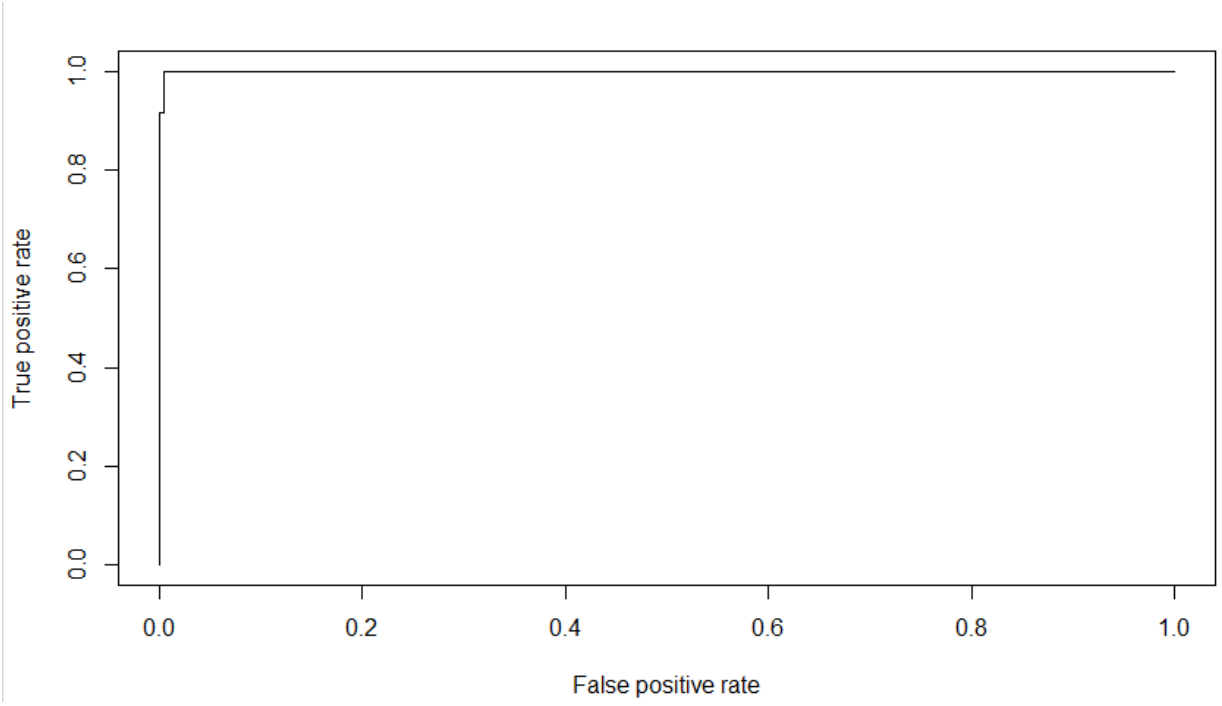

```
> table(datos[-training.ids,]
+       dnn = c("Actual","Pre
       Predicho
Actual   0   1
     0 195  30
     1  18  57
> |
```



Universidad Rafael Landívar
Facultad de Ingeniería
Análisis de Datos
Msc. Ing. Pablo Orellana

Random Forest

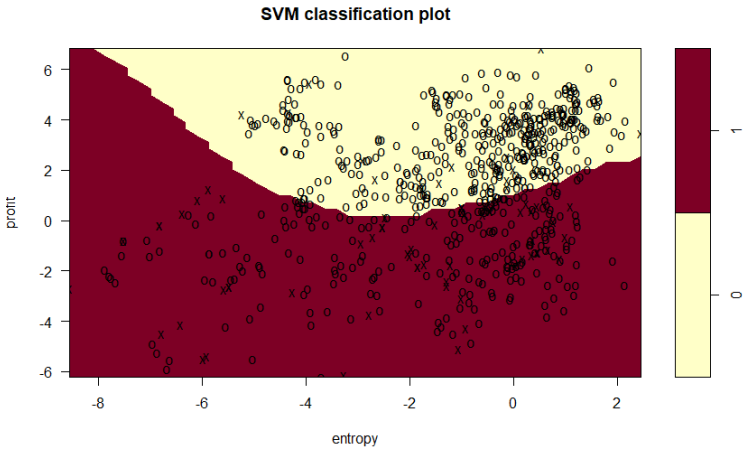
```
> table(datos2[-training.ids2,"online"],pred2,dnn = c("Actual","Predicho"))
      Predicho
Actual   0   1
     0 227   1
     1   2  69
> |
```

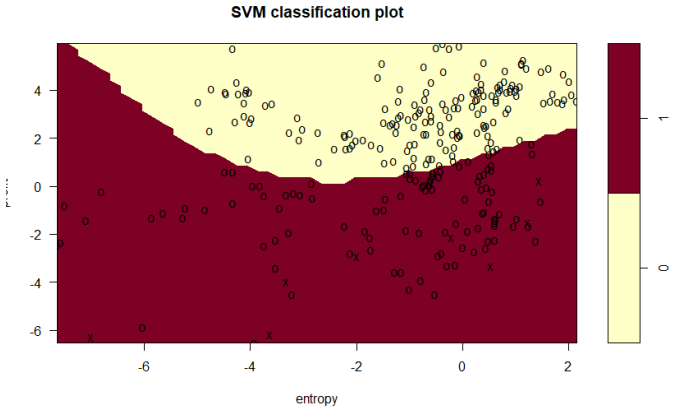


Universidad Rafael Landívar
Facultad de Ingeniería
Análisis de Datos
Msc. Ing. Pablo Orellana

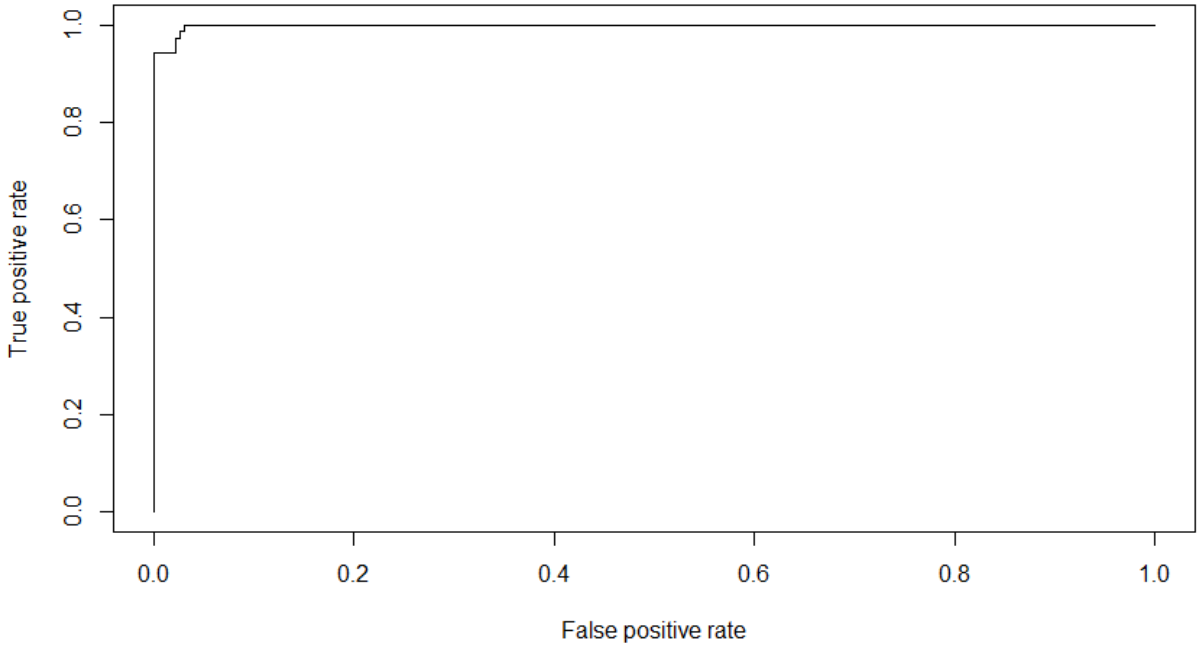
SVM

```
> table(datos3[-ti
       Predicho
Actual   0   1
     0 228   0
     1   0  71
> |
```





Universidad Rafael Landívar
Facultad de Ingeniería
Análisis de Datos
Msc. Ing. Pablo Orellana



Serie 4 (35 puntos)

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

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

***Ver documento adjunto

Universidad Rafael Landívar
Facultad de Ingeniería
Análisis de Datos
Msc. Ing. Pablo Orellana

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:

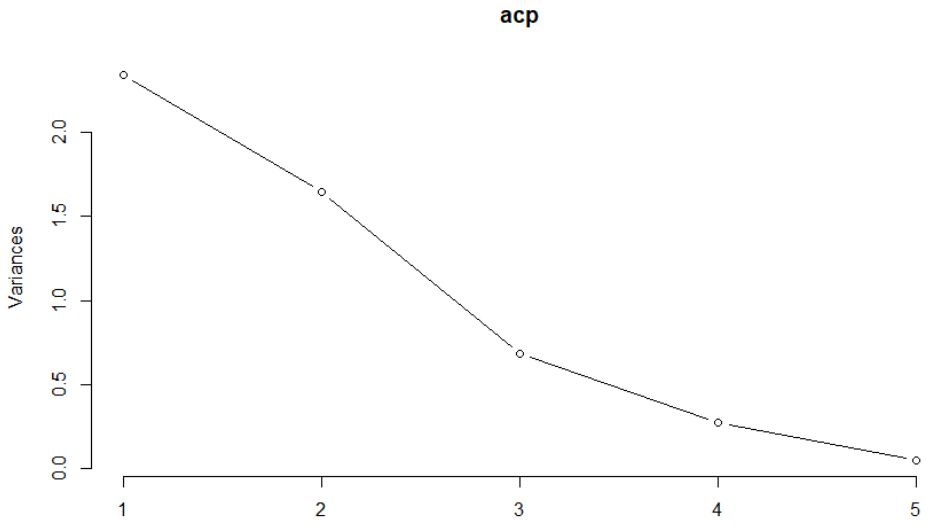
- **Cuantos componentes principales utilizará para su análisis**

  Se utilizarán 2 componentes: PC1 y PC2.

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

  0.47 y 0.8 Respectivamente

```
Importance of components:
                          PC1    PC2    PC3     PC4     PC5
Standard deviation     1.5312 1.2842 0.8261 0.52464 0.21989
Proportion of Variance 0.4689 0.3298 0.1365 0.05505 0.00967
Cumulative Proportion  0.4689 0.7988 0.9353 0.99033 1.00000
> 
```



Universidad Rafael Landívar
Facultad de Ingeniería
Análisis de Datos
Msc. Ing. Pablo Orellana