Primer Parcial (30pts)

Programacion Estadística II

Lic. Alvaro Chirino Gutierrez

11/7/2020

Instrucciones

- Duración: 2 horas. 18:00 a 20:00
- Forma de entrega, incluir solo las respuestas en archivos separados, el nombre del archivo debe ser: respuesta_1, respuesta_2, etc.
- La entrega al correo achirino.stat@gmail.com máximo hasta las 20:00, pasado el tiempo no se tomara en cuenta el envío.
- Incluir en el correo su nombre completo.

Pregunta 1 (10 pts)

Usando la encuesta a hogares 2018, realice de forma secuencial lo siguiente:

- Seleccione al jefe/jefa del hogar de hogares que tengan al menos 1 niño/a de 5 años o menos
- Para la variable ingreso laboral (en logaritmo natural) del jefe del hogar construya un modelo lineal con las siguientes variables dependientes:
  - edad
  - sexo
  - número de miembros
  - número de personas de 60 años o más
  - años de educación
  - hogar con acceso a electricidad
  - servicio sanitario privado
  - hogar pobre/no pobre
- Realice los siguientes pasos sobre el modelo
  - Según la naturaleza de las variables defina el modelo lineal y elimine a las no significativas ($p < 0.05$)
  - Realice el test de normalidad sobre los errores y comente
  - Realice un test de multicolinealidad y comente
  - Realice un test de Homocedasticidad y comente
  - Incluya la interacción edad con número de miembros
  - Incluya los polinomios hasta el grado 3 para la variable edad

Pregunta 2 (10 pts)

Usando la ENDSA para el 2008 para el corte de edad de 20 a 40 años, para la variable fumar; elabore un modelo de clasificación usando los métodos:

- Logístico
- Árbol de clasificación
- Naive Bayes

Considere al menos 10 covariables para la elaboración de los métodos. Comente el mejor método de clasificación, justifique su respuesta.

Pregunta 3 (10 pts)

Seleccione solo una de las siguientes fuentes de información:

1. Usando la API de twitter genere una base de datos de los 1000 últimos post de la palabra "coronavirus", sin retweets, en español.
2. Colección de 3 tesis en pdf del enlace https://repositorio.umsa.bo/handle/123456789/7292

A partir de esta base realice de forma secuencial:

- Establezca el Corpus
- Ponga las palabras en mayúsculas
- Elimine los números y la puntuación
- Elimine los stopwords en español e incluya 3 palabras definidas por usted para su eliminación
- Realice una nube de palabras y comente los hallazgos
- Realice el análisis de sentimiento y grafique 2 figuras; una para las ocho emociones y otra para los sentimientos positivos y negativos

Pregunta opcional (5 pts)

Usando la librería syuzhet incorpore las siguientes 4 palabras en el léxico nrc:

- racista, negative, 1
- patético, anger,1
- roban, anger,1
- lacra, anger,1

Adapte la función para que el análisis de sentimiento incluya estas palabras. Verifique el funcionamiento con la frase:

"Esa persona es racista, patético y una lacra. otros roban a las personas"

Ejemplo en R

```
frase<-"Esa persona es racista, patético y una lacra. otros roban a las personas"
#resultado actual
library(syuzhet)
get_nrc_sentiment(frase,language = "spanish")
```

```
##   anger anticipation disgust fear joy sadness surprise trust negative positive
## 1     0            0       0    0   0       0        0     0        0        0
```

```
#resultado esperado
ejemplo(frase,language = "spanish")
```

```
##   anger anticipation disgust fear joy sadness surprise trust negative positive
## 1     2            0       0    0   0       0        0     0        1        0
```