Segunda Clase de Análisis de Datos

Prof: Boris Panes
Universidad Del Desarrollo

Agosto 24, 2024

Contenido

Profundización en los temas de limpieza y transformación de datos

- Proceso experimental
- Revision de elementos de cada proceso
- Ejemplo práctico
  - Aspectos positivos y negativos de cada herramienta

Coordinación de proyectos T1

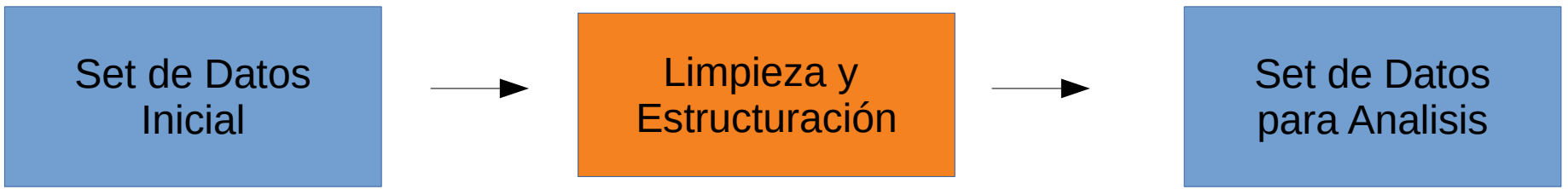
- Discusion sobre bases de datos disponibles
- Detalles del proceso de investigación
- Definición del método de evaluación

La Ciencia de Datos es una Ciencia Experimental

Experimetos en ciencia de datos

Consideremos una estructura de datos estandar, donde las filas representan eventos independientes (ocurrencias) y las columnas los datos asociados a cada evento (propiedades)

El proceso de observación y registro por lo general involucra el filtro y manipulación de los datos generados por los eventos. Por lo tanto es esperable que este proceso contenga inconsistencias y errores. Es una suposición conservadora sobre cualquier set de datos que se busque estudiar



Selección de datos

Para esta clase buscaremos un set de datos que nos permita explorar los conceptos relacionados con la limpieza y estructuración de datos

Volvamos a recordar los elementos claves relacionados con la Calidad de los Datos

Exploración y Limpieza de datos

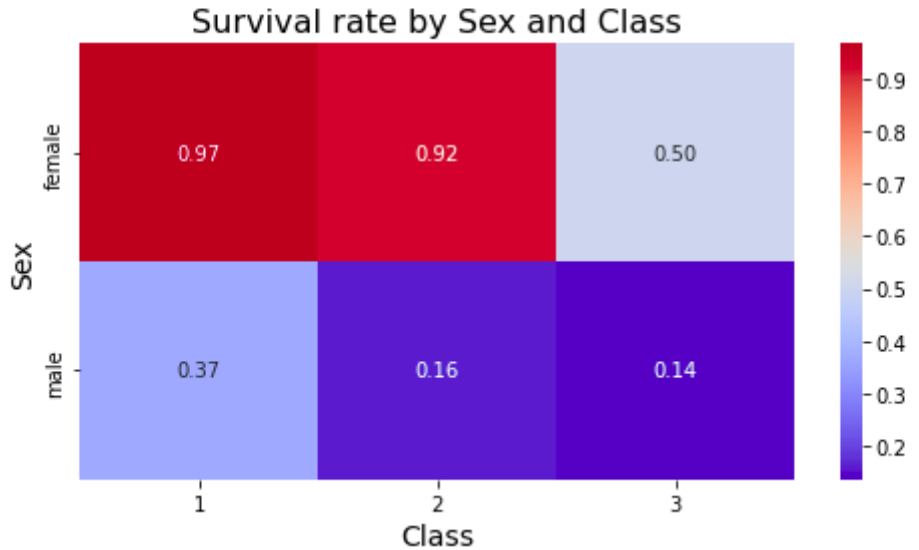
Revisión del contenido de las columnas y filas
Formato de los datos: fechas, codigos, identificadores
Datos duplicados: filas y/o columnas repetidas
Completitud de los datos: contenido invalido, como por ejemplo NaN

Transformación de los datos

Ingenieria de caracteristicas: reemplazo de valores, suma de columnas
Estadarización: distribución normal estandar
Normalización: restricción del intervalo
ETL: extract, transform and load (combinación de multiples fuentes de datos)

Que tal si juntamos las variables mas representativas hasta el momento (Sex y PClass)

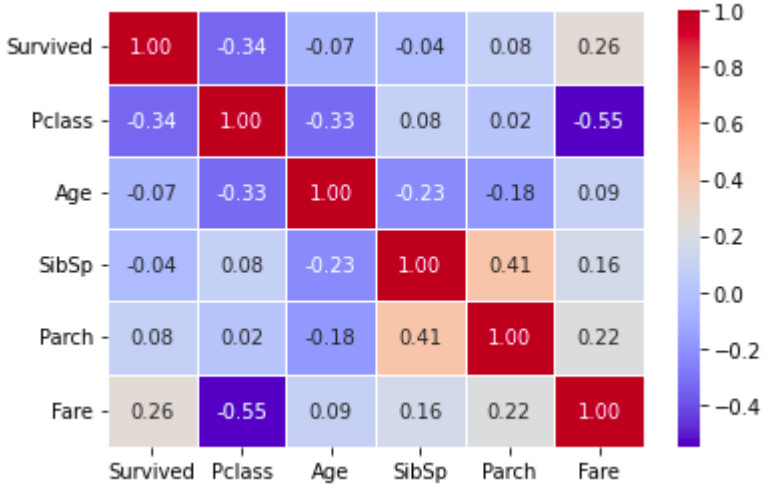
```
In [78]: sex_class_survival = df.groupby(['Sex','Pclass'])['Survived'].mean().unstack()
         # plot
         plt.figure(figsize=(8,4))
         sns.heatmap(sex_class_survival, annot=True, fmt='.2f', cmap='coolwarm')
         plt.title('Survival rate by Sex and Class', fontsize=16)
         plt.ylabel('Sex', fontsize= 14)
         plt.xlabel('Class',fontsize=14)
         plt.show()
```



Podemos ver que la combinación entre PClass=1 y Sex=Female muestra la tasa mas alta de sobrevivencia con un 97%

Por otro lado, PClass=3 y Sex=Male obtiene la menor tasa de sobrevivencia con 14%

Pasando de dos columnas a una sola



Existe una moderada correlación positiva entre las variables SibSp y Parch, donde

SibSp = hermanos y parejas abordo
Parch = padres e hijos

SibSp + Parch = Relatives

Este proceso en general se denomina como ingeniería de datos, dado que se esta creando un nuevo tipo de característica a partir de los tipos nativos

```
In [28]: df["Familysize"] = df["SibSp"] + df['Parch']
```

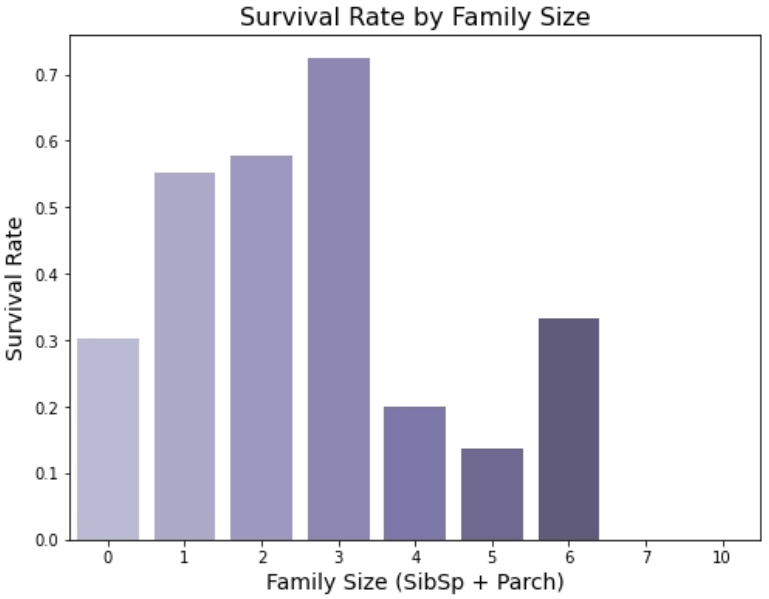
```
In [29]: family_surr = df.groupby('Familysize')['Survived'].mean()
```

```
In [30]: family_surr
```

```
Out[30]: Familysize
         0     0.303538
         1     0.552795
         2     0.578431
         3     0.724138
         4     0.200000
         5     0.136364
         6     0.333333
         7     0.000000
         10    0.000000
         Name: Survived, dtype: float64
```

```
In [31]: #plot
         plt.figure(figsize=(8,6))
         sns.barplot(x=family_surr.index, y=family_surr.values,
                     palette='Purples_d')
         plt.title('Survival Rate by Family Size', fontsize=16)
         plt.xlabel('Family Size (SibSp + Parch)', fontsize=14)
         plt.ylabel('Survival Rate', fontsize=14)
         plt.show()
```

Apoyo más facilidad de movimiento:
Los pasajeros con familias medianas presentan una mayor probabilidad de sobrevivir



Coordinación de proyectos

Requerimientos mas especificos

Materiales y formato de entrega:

Set de datos en formato csv (comma separated values) plano, multicolumna,
Notebook escrito en python donde se carga el set de datos y se interactua
Presentación en PDF, 10-15 láminas
Video describiendo el trabajo **(entrega por correo o CANVAS)**

Contenido del Video:

Introduccion al trabajo por parte del presentador (1 min)
Presentacion de los integrantes y resumen de su contribución (5 minutos)
Presentacion del set de datos (1 min)
Definicion del objetivo general asociado al set de datos (1 min)
Alcance del analisis desarrollado y relación con el objetivo final (2 min)
Descripcion del notebook para justificar su argumento (5 min)

Tiempo total aproximado: 15-20 minutos

Requerimientos mas especificos

Alcance del analisis desarrollado y relación con el objetivo final (2 min)

En este notebook exploramos ciertas variables X, Y y Z del set de datos
Estas variables pueden ser utiles para entender A, B o C sobre el objetivo general del desafio asociado al set de datos.
El analisis realizado incluye un revision del contenido de X, Y o Z, busqueda de valores nulos, duplicados, formato, transformaciones, etc

Descripcion del notebook para justificar su argumento (5 min)

Seguimiento de 2 a 4 laminas con extractos del notebook que validen su analisis anunciado en el punto anterior
Conclusion

Fecha y hora de entrega: 07 de Septiembre de 2024 a las 23:259 hrs

Grupos

Notebooks potables

https://www.kaggle.com/code

ashydv/housing-price-prediction-linear-regression
data13/predicting-house-prices-with-linear-regression
nakulmalik/house-prices-linear-regression

Regresión lineal
Normalización
Variables mudas
Parece simple y es publico (investigar al autor, seguir con ojo critico)

sukhyun5/steel-plate-faults-data-analysis-with-r

set de datos entendible, pero es un problema de clasificacion y esta bastante limpio

rautaishwarya/data-cleaning-and-price-prediction

el notebook se ve muy bueno en cuanto a formato de datos

qusaybtoush1990/wine-quality

notebook con muchos votos. un poco de **formateo y exploración**

Ejercicio práctico modo proyecto

Preguntas con alternativas

1) Cuantas combinaciones de 8 grupos con 4 estudiantes cada uno se pueden hacer en un curso con N=36 estudiantes

1- Calcular la cantidad de grupos de 4 personas se pueden hacer con N=36

2- Luego calcular cuantos grupos de 8 bloques podemos hacer con el numero total de grupos disponibles