**PROYECTO FINAL DE DATASCIENCE IN HEALTH CARE**

**Antes de empezar**

1- Elija un conjunto de datos relacionados con el sector salud. Idealmente, el conjunto de datos debería tener alrededor de ~ 200 filas (Observaciones) y debería tener variables (predictores) tanto categóricas como numéricas.

Fuentes potenciales de datos: TidyTuesday: <https://github.com/rfordatascience/tidytuesday>

* KAGGLE. <https://www.kaggle.com/general/168211>
* <https://medium.com/@ODSC/15-open-datasets-for-healthcare-830b19980d9>

2- Debe referenciar la fuente donde extrajo:

Javier Fontalvo:

La base de datos fue obtenida de una fuente externa, a través de un contacto en el sector salud.

3- Tenga en cuenta que la mayoría de estos son archivos .csv. Escriba el código para cargar los archivos desde csv para cada uno de los conjuntos de datos y una breve descripción de las variables, o puede cargar el .csv archivo en su carpeta de datos.

**Trabajando juntos**

Los Integrantes del grupo son Javier Fontalvo, Luis Parra, Diego reyes.

Cada integrante especificara que hizo y cómo fue su aporte en los respectivos espacios de trabajo del documento.

**Defina su pregunta de investigación (10 Puntos)**

4 - Defina su pregunta de investigación a continuación. ¿Qué te interesa de los datos? ¿Cuál es una pregunta específica que desea averiguar sobre los datos? (10 Puntos).

**Javier Fontalvo:**

**de la base de datos escogida, se puede ver las diferentes remisiones ocurridas en un periodo de 3 meses del año 2020 en un hospital de primer nivel en la zona del Urabá antioqueño, donde se puede ver a que ciudades fueron remitidos los pacientes, con que diagnósticos, a que eps pertenecen, etc.**

**La pregunta que nos gustaría averiguar seria ¿Qué eps tiene más a tendencia a que la remisión se aceptada?**

**librerías**

**5 -** Dada su pregunta, ¿cuál es su expectativa sobre los datos?, es decir, ¿qué conclusiones espera obtener del análisis de sus datos? (10 Puntos)

Javier Fontalvo:

se espera ver la cantidad de remisiones que tuvo cada una de las eps en el año 2020.

**Carga y preprocesamiento los datos (10 Puntos)**

Luis Parra Pacheco

6- Cargue los datos a continuación y utilice dplyr::glimpse() o skimr::skim()sobre los datos. Debe cargar el archivo de datos en el directorio. Limpieza de los datos.

#Luis Parra- Se cargan los datos del dataset montado en la carpeta local del usuario

Remisiones\_del\_hospital <- read\_excel("C:/Users/pluis/OneDrive - UNIVERSIDAD AUTONOMA DEL CARIBE/Remisiones del hospital.xlsx")

View(Remisiones\_del\_hospital)

#Se transforma a un dataframe para manejar los datos.

Remisiones\_del\_hospital <- as.data.frame(Remisiones\_del\_hospital)

**Transformando los datos (15 puntos)**

Luis Parra Pacheco

7- Como la pregunta de investigación es ver la cantidad de remisiones que tuvo cada eps, lo primero que se hace es crear un nuevo objeto que almacena las eps que son de tipo String, luego convertirlo a factor, reemplazar los valores del dataframe original por los ya procesados y aplicar un summary a la columna de eps y muestra las cantidades que tiene cada eps.

**Visualizar y resumir los datos (25 puntos)**

Luis Parra Pacheco

8- Se utilizo Barplot para graficar los factores de la columna E.P.S donde se visualizó un claro ganador.

9 ¿Cuáles son sus hallazgos sobre el resumen? ¿Son lo que esperabas? Usar las métricas para evaluar los modelos, por ejemplo: matriz confusión, sensibilidad … Curva ROC, etc

Diego Reyes Serpa

**Conclusiones (40 puntos)**

10- Resuma su pregunta de investigación y sus hallazgos a continuación.

Diego Reyes Serpa:

Nuestra pregunta de investigación fue simple y certera de una base de datos de remisiones medicas en un hospital de la zona de uraba, queríamos determinar y hallar cual era la EPS con mayor cantidad de remisiones durante el tiempo de estudio de la muestra. Los resultados hallados muestran el predominio de dos Eps Savia Salud y Nueva Eps, esta última fue formada hace poco tiempo, pero incluso así presenta una gran cantidad de miembros; otro hallazgo que resulto fuera de lo esperado es que no hubo un gran número de remisiones por parte del Sisbén, esperábamos encontrar mayor cantidad de remisiones esto debido a que esta cerca de un área rural y la gran mayoría de esta población están suscritos al Sisbén pero resultaron siendo muy escasas las remisiones por parte del Sisbén

11- ¿Son sus hallazgos los que esperaba? ¿Por qué o por qué no?

Diego Reyes Serpa:

Los hallazgos que encontramos en nuestro reporte de remisiones de las distintas EPS fueron satisfactorios además de ser esperados debido a que en el lugar que fueron seleccionadas las muestras la EPS que más acompañamiento y planes promocionales para las personas es “Savia Salud” por lo tanto se esperaba que esta fuera la que presentar mayor cantidad de remisiones al hospital seleccionado, si las muestras hubieran sido tomadas en otro lugar se esperaría un resultado diferente pero teniendo en cuenta lo antes mencionado estamos de acuerdo que era un resultado esperado además de predecible

**Git Hub (10 puntos)**

Subir todo el código y el reporte de R-markdown en GitHub

Compartir el enlace (colocarlo público)

Javier fontalvo:

<https://github.com/jfontal/parcial-final-ciencia-de-datos-2021-02>

**### Solucion Parcial 3: Data science for Healthcare**

**\*Introducción\***

**Los hospitales diariamente reciben muchos pacientes los cuales no pueden ser tratados en la sede, por lo cual se les diagnostica, y remiten a otros hospitales, las bases de datos con las que llevan control poseen datos muy importantes para la estadística de dicho hospital y en este caso hipotético se necesita saber cuál es la E.P.S que con mayor frecuencia sus pacientes necesitan remisión.**

**### \*Punto 1:\* Referenciar la fuente donde extrajo el dataset, Defina su pregunta de investigación y ¿Qué conclusiones espera obtener del análisis de sus datos? . \*\*(10 Puntos)\*\***

**Javier Fontalvo: La base de datos fue obtenida de una fuente externa, a través de un contacto en el sector salud, de la base de datos escogida, se puede ver las diferentes remisiones ocurridas a través del año 2020 en un hospital de primer nivel en la zona del Urabá antioqueño, donde se puede ver a que ciudades fueron remitidos los pacientes, con que diagnósticos, a que eps pertenecen, etc.**

**Se espera ver la cantidad de remisiones que tuvo cada una de las eps en el año 2020.**

**### La pregunta que nos gustaría averiguar seria \*¿Qué eps tiene más a tendencia a que la remisión se aceptada?\***

**Cargar librerías**

**```{r, warning= FALSE, message = FALSE }**

**library(readxl)**

**library(dplyr)**

**library(ggplot2)**

**library(hrbrthemes)**

**```**

**### \*Punto 2:\* Carga y preprocesamiento de los datos. \*\*(10 Puntos)\*\***

**```{r, heart, warning= FALSE, message = FALSE }**

**Remisiones\_del\_hospital <- read\_excel("C:/Users/pluis/OneDrive - UNIVERSIDAD AUTONOMA DEL CARIBE/Remisiones del hospital.xlsx")**

**View(Remisiones\_del\_hospital)**

**Remisiones\_del\_hospital <- as.data.frame(Remisiones\_del\_hospital)**

**glimpse(Remisiones\_del\_hospital)**

**str(Remisiones\_del\_hospital)**

**summary(Remisiones\_del\_hospital)**

**```**

**### \*Punto 3:\* Transformando los datos. \*\*(15 puntos)\*\***

**Luis Parra Pacheco: Como la pregunta de investigación es ver la cantidad de remisiones que tuvo cada eps, lo primero que se hace es crear un nuevo objeto que almacena las eps que son de tipo String, luego convertirlo a factor, reemplazar los valores del dataframe original por los ya procesados y aplicar un summary a la columna de eps y muestra las cantidades que tiene cada eps.**

**```{r, heart, warning= FALSE, message = FALSE }**

**glimpse(Remisiones\_del\_hospital)**

**str(Remisiones\_del\_hospital)**

**summary(Remisiones\_del\_hospital)**

**E.P.S <- Remisiones\_del\_hospital$E.P.S**

**E.P.S <- as.factor(E.P.S)**

**Remisiones\_del\_hospital$E.P.S <- E.P.S**

**summary(Remisiones\_del\_hospital$E.P.S)**

**```**

**### \*Punto 4:\* Visualizar y resumir los datos. \*\*(25 puntos)\*\***

**Luis Parra Pacheco: Se utilizo Barplot para graficar los factores de la columna E.P.S donde se visualizó un claro ganador.**

**```{r, heart, warning= FALSE, message = FALSE }**

**barplot(table(Remisiones\_del\_hospital$E.P.S),**

**main="Frecuencia de Remisiones",**

**xlab="E.P.S",**

**ylab="Remisiones",**

**col="blue",**

**density=20)**

**```**

**### \*Punto 5:\* Conclusiones. \*\*(40 puntos)\*\***

**Resuma su pregunta de investigación y sus hallazgos a continuación.**

**Diego Reyes Serpa: Nuestra pregunta de investigación fue simple y certera de una base de datos de remisiones medicas en un hospital de montería queríamos determinar y hallar cual era la EPS con mayor cantidad de remisiones durante el tiempo de estudio de la muestra. Los resultados hallados muestran el predominio de dos Eps Savia Salud y Nueva Eps, esta última fue formada hace poco tiempo, pero incluso así presenta una gran cantidad de miembros; otro hallazgo que resulto fuera de lo esperado es que no hubo un gran número de remisiones por parte del Sisbén, esperábamos encontrar mayor cantidad de remisiones esto debido a que esta cerca de un área rural y la gran mayoría de esta población están suscritos al Sisbén pero resultaron siendo muy escasas las remisiones por parte del Sisbén**

**Los hallazgos que encontramos en nuestro reporte de remisiones de las distintas EPS fueron satisfactorios además de ser esperados debido a que en el lugar que fueron seleccionadas las muestras la EPS que más acompañamiento y planes promocionales para las personas es “Savia Salud” por lo tanto se esperaba que esta fuera la que presentar mayor cantidad de remisiones al hospital seleccionado, si las muestras hubieran sido tomadas en otro lugar se esperaría un resultado diferente pero teniendo en cuenta lo antes mencionado estamos de acuerdo que era un resultado esperado además de predecible.**