**Programa Ciencia de Datos – FUNDATEC**

**Módulo 4 – Curso Big Data**

**Proyecto**

**Entregable #1**

Sebastián Porras

Setiembre, 2022

De forma detallada, deberá abordarse lo siguiente:

● Fuentes de datos analizadas. Los estudiantes deben documentar qué fuentes de datos analizaron. Deberán escoger al menos dos fuentes de datos que se puedan cruzar exitosamente, para obtener un conjunto de datos de mayor riqueza de información. Los estudiantes deberán argumentar por qué realizaron la selección final. A manera de sugerencia, los estudiantes pueden tratar de utilizar datos del INEC, Programa Estado de la Nación, Ministerios de Gobierno (e.g. Economía, Educación), por nombrar algunos. El requerimiento estricto, eso si, es que se puedan cruzar unos con los otros.

● Descripción detallada de los datos. Solamente para los datos escogidos, deberán describir cada uno de los atributos contenidos. También deberá explicarse cómo se une un conjunto de datos a otro (e.g. por número de cédula). Los estudiantes podrán utilizar las técnicas ya aprendidas para mejorar el entendimiento de los lectores, por ejemplo, estadística descriptiva, distribuciones, etc.

● Objetivo predictivo. Deberá explicarse en detalle qué atributo de los datos se utilizarán como variable objetivo del modelo de aprendizaje automático. Esto servirá como el planteamiento del objetivo de investigación que se plantea, antes de iniciar la realización del proyecto.

Fuentes de Datos

**US\_Accidents\_Dec21\_updated**: Dataset extraido de Kaggle, que describe los accidentes que ocurrieron en 2021 en Estados Unidos

Columnas a utilizar

1. Severity: La severidad del accidente.
2. Start\_Time: Fecha y hora en la que el accidente se dio.
3. End\_Time: Fecha y hora en que el accidente se dio por terminado .
4. City: Ciudad en la que se dio el accidente.
5. County: Condado en el que se dio el accidente.
6. State: Estado en el que ocurrió el accidente.
7. Temperature(F) : Temperatura captada cuando se dio el accidente.
8. Wind\_Chill(F): Sensación térmica en el momento del accidente.
9. Humidity(%): La humedad en el momento del accidente.
10. Pressure(in): Presión atmosférica en el momento del accidente.
11. Visibility(mi) : Visibilidad reportada.
12. Zipcode: Código postal

Distribución de los valores de severidad

Imagen que contiene Gráfico de barras

Descripción generada automáticamente

Representación grafica de la visibilidad cuando ocurrieron los accidentes

Gráfico

Descripción generada automáticamente

**WeatherEvents\_Jan2016-Dec2021:** Dataset extraido de Kaggle, que describe los eventos meteorológicos desde Enero del 2016 hasta Diciembre del 2021

Columnas a utilizar

1. Type: Describe el tipo de precipitación (Nieve, tormenta, neblina, etc)
2. Severity: Describe que tan fuerte fue esta precipitación
3. StartTime(UTC): Fecha y hora de la precipitación
4. EndTime(UTC): Fecha y hora de la precipitación
5. Precipitation(in): Cantidad en pulgadas de lo que precipitó
6. City: Ciudad en donde se dio el fenómeno
7. County: Condado en donde se dio el fenómeno
8. State: Estado en donde se dio el fenómeno
9. ZipCode: Código postal

Objetivo predictivo: Con este trabajo intento predecir, dado un conjunto de caracteristicas del clima, que tan severo podría ser un accidente en caso de que se dé.

Distribución de los valores de la columna Type

Gráfico

Descripción generada automáticamente con confianza media

Distribución de los valores de la columna Severity

Gráfico

Descripción generada automáticamente