UNIVERSIDAD AUTÓNOMA DE QUERÉTARO
FACULTAD DE INGENIERÍA
DIVISIÓN DE INVESTIGACIÓN Y POSGRADO

Machine learning: Práctica 1. Análisis estadístico y visualización de datos.

Alumno:
Ing. Enrique Mena Camilo

Profesor:
Dr. Marco Antonio Aceves Fernández

Febrero 2023

Índice

1. Objetivo

Desarrollar habilidades en el análisis estadístico y visualización de datos en el lenguaje de programación Python, mediante la construcción de diversos tipos de gráficos y la obtención de métricas como:

- Máximo.
- Mínimo.
- Desviación estándar.
- Mediana.
- Moda.
- Datos faltantes.
- Balance de clases.
- Tipo de distribución.
- Valores atípicos.

2. Introducción

El componente principal de todo modelo de inteligencia artificial son los datos, y mientras mejores datos tengamos al alcance mejores serán los resultados de nuestro modelo.

Parte de contar con buenos datos consiste en conocer los mismos, determinar sus rangos, las relaciones entre sus atributos, la distribución de cada uno de ellos, y algunas métricas de estadistica descriptiva de cada atributo.

Se espera que el desarrollo de esta práctica aporte experiencia y desarrolle habilidades en el proceso de obtención de métricas de estadistica descriptiva y visualización de datos.

3. Marco teórico

3.1. Instancias

En general, los sistemas tienen entradas y salidas. Específicamente los conjuntos de datos entrada y salida pueden ser instancias, atributos y características. Si aterrizamos estos conceptos en una base de datos, las instancias serían las filas. [1]

3.2. Atributos

Los atributos o registros, corresponden a una instancia particular, en la matriz de la base de datos son las columnas. Existen distintos tipos de atributos: numéricos, nominales, categóricos ordinales, entre otros. [1]

3.3. Estadística de atributos

Cuando se trabaja con datos, es importante saber el tipo de operaciones, cálculos y medidas que nos permiten conocer la tendencia de su comportamiento. A continuación, se describen algunas medidas importantes:

- **Máximo**: Corresponde al valor más grande contenido en un conjunto de datos o en un atributo.
- **Mínimo**: Corresponde al valor más pequeño contenido en un conjunto de datos o en un atributo.
- **Desviación estándar**: Se define como una medida de la dispersión de un conjunto de datos. A mayor desviación estándar, mayor dispersión de los datos. Su expresión matemática es la siguiente:

$$s = \sqrt{\frac{\Sigma(x_i - \bar{x})^2}{n-1}}$$

- **Media**: Es una medida de la tendencia central de los datos. Es la suma de los valores de los datos entre la cantidad de datos.
- **Moda**: Dentro de los atributos corresponde al valor que más se repite dentro de los atributos.

3.4. Tipo de distribución

El histograma permite conocer el tipo de distribución de los datos con los que se trabajará, existen distintos tipo de distribución, dentro de los cuales se puede destacar [1]:

- **Uniforme**: Mostrada en Figura 1.
- **Normal (unimodal)**: Mostrada en Figura 2.
- **Unimodal sesgada izquierda**: Mostrada en Figura 3.
- **Unimodal sesgada derecha**: Mostrada en Figura 4.
- **Multimodal**: Mostrada en Figura 5.

Figura 1: Ejemplo de distribución uniforme.

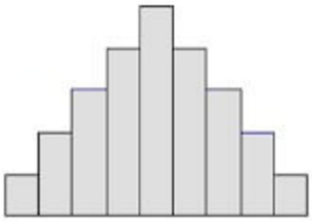


Figura 2: Ejemplo de distribución normal.

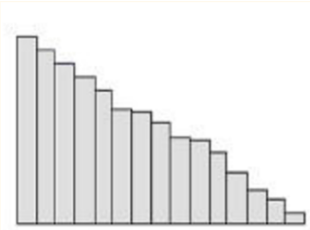


Figura 3: Ejemplo de distribución sesgada a la izquierda.

3.5. Relaciones entre atributos

- **Cuartiles**: Se definen como los valores que dividen un conjunto de datos en cuatro subconjuntos que poseen alrededor del mismo número de observaciones. El total de los datos corresponde al 100 %, y este se divide en 25 %, 50 %, 75 % y 100 %. [2]

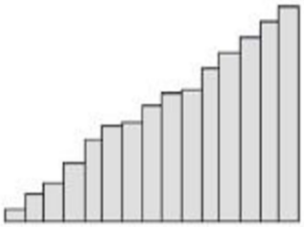


Figura 4: Ejemplo de distribución sesgada a la derecha.

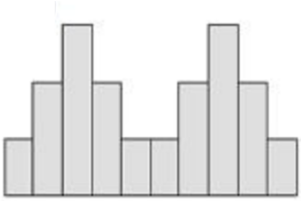


Figura 5: Ejemplo de distribución multimodal.

- **Varianza**: Es la medida de dispersión de los datos con respecto a la media de los mismos. [3]
- **Covarianza**: Mide la variabilidad entre dos variables. Su valor será positivo si las variables si la relación entre ellas es lineal y van en la misma dirección, en cambio, será negativa si su relación es inversa. [4]

3.6. Datos atípicos en atributos

Resulta importante conocer la calidad de los datos con los que se trabajará, y los problemas que este pueda presentar por su naturaleza misma. Existen problemas como valores faltantes, problemas con la cardinalidad irregular y valores atípicos. [1]

Particularmente los valores atípicos son aquellos que están muy alejados de la tendencia central. Pueden ser causa de errores cuando el registro de los datos se genera de manera manual o puede suceder cuando realmente un valor está muy alejado del resto, debido al tipo de información que se está registrando. [1]

4. Materiales y métodos

Para el desarrollo de la práctica se utilizó el lenguaje de programación Python en su versión 3.10, el cual fue utilizado para desarrollar una Jupyter Notebook con los requerimientos de la práctica.

El conjunto de datos a utilizar proviene de un repositorio de acceso público de la Secretaría de Salud del Gobierno de México, y consiste en datos tabulares anonimizados con información relevante a casos de COVID-19 en México.

Para la obtención de los datos estadísticos se utilizaron métodos de los paquetes *Numpy* y *Statistics* de Python, los cuales se consumieron dentro de una clase diseñada para agilizar el proceso de obtención de datos estadisticos.

Para la visualización de los datos se utilizaron dos paquetes de Python especializados en la visualización de datos interactiva, *Bokeh* y *Pygal*, los cuales pueden ser instalados en el interprete de Python mediante el gestor de paquetes PIP con los comandos:

```
pip install Pygal
pip install bokeh
```

Para el caso especial del paquete *Bokeh* se requiere de un paquete adicional, el cual se puede instalar de igual forma con el gestor de paquetes PIP:

```
pip install selenium
```

5. Pseudocódigo

El proceso a seguir para el desarrollo de esta práctica está definido por le pseudocódigo mostrado a continuación.

```
Proceso AnalisisConjuntoDatos
        Leer datos;
        Eliminar atributos innecesarios;
        Dar formato a datos;
        Obtener estadisticos de datos;
        Generar visualizaciones de datos;
FinProceso
```

6. Resultados

6.1. Análisis estadístico

La Tabla 1 muestra los resultados obtenidos para el análisis estadistico del conjunto de datos, donde se agrupan los resultados por atributo. En dichos resultados, hay información que no es posible obtener debido a la naturaleza del atributo, dichos datos son representados con *NA*.

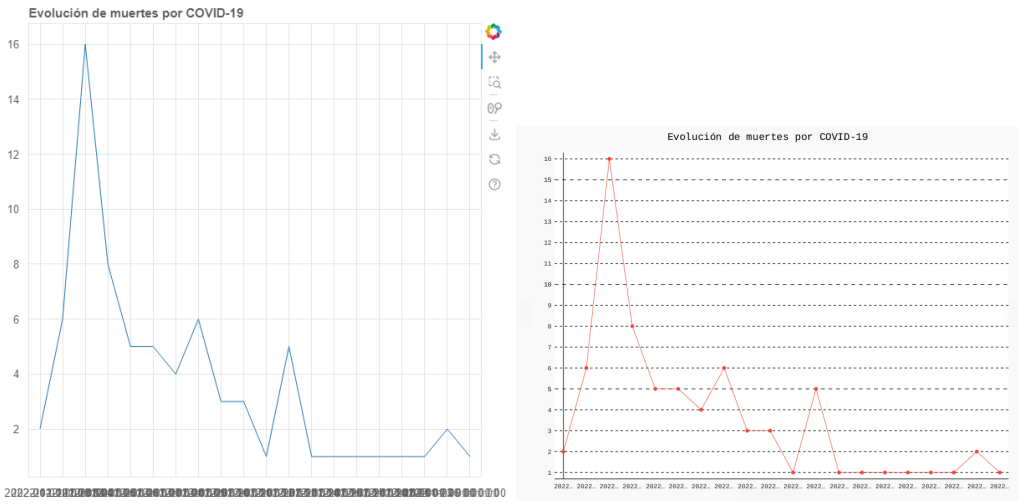
Atributo	Tipo de dato	Mínimo	Máximo	Moda	Datos faltantes	Mediana	Desviación estándar	Promedio
ORIGEN	Entero	1	2	2	0	2	0	1
SECTOR	Entero	1	13	12	0	12	0	9
ENTIDAD_UM	Entero	1	32	9	0	9	0	13
SEXO	Entero	1	2	1	0	1	0	1
ENTIDAD_NAC	Entero	1	99	9	81	9	0	14
ENTIDAD_RES	Entero	1	32	9	0	9	0	13
MUNICIPIO_RES	Entero	1	530	5	23	13	0	25
TIPO_PACIENTE	Entero	1	2	1	0	1	0	1
FECHA_INGRESO	Fecha	2022-01-01	2023-02-03	2022-01-03	0	*NA*	*NA*	*NA*
FECHA_SINTOMAS	Fecha	2022-01-01	2023-02-03	2022-01-01	0	*NA*	*NA*	*NA*
FECHA_DEF	Fecha	2022-01-01	2022-12-31	2022-12-31	0	*NA*	*NA*	*NA*
INTUBADO	Entero	1	99	97	2	97	0	91
NEUMONIA	Entero	1	2	2	0	2	0	1
EDAD	Entero	0	9	30	2	36	0	37
NACIONALIDAD	Entero	1	2	1	0	1	0	1
EMBARAZO	Entero	1	98	2	46	2	0	44
HABLA_LENGUA_INDIG	Entero	1	99	2	648	2	0	8
INDIGENA	Entero	1	99	2	635	2	0	8
DIABETES	Entero	1	98	2	59	2	0	2
EPOC	Entero	1	98	2	58	2	0	2
ASMA	Entero	1	98	2	55	2	0	2
INMUSUPR	Entero	1	98	2	56	2	0	2
HIPERTENSION	Entero	1	98	2	54	2	0	2
OTRA_COM	Entero	1	98	2	82	2	0	2
CARDIOVASCULAR	Entero	1	98	2	54	2	0	2
RENAL_CRONICA	Entero	1	98	2	57	2	0	2
TABAQUISMO	Entero	1	98	2	54	2	0	2
OTRO_CASO	Entero	1	99	2	165	2	0	3
TOMA_MUESTRA_LAB	Entero	1	2	2	0	2	0	1
RESULTADO_LAB	Entero	1	97	97	0	97	0	73
TOMA_MUESTRA_ANTIGENO	Entero	1	2	1	0	0	0	1
RESULTADO_ANTIGENO	Entero	1	97	2	0	2	0	16
CLASIFICACION_FINAL	Entero	1	7	7	0	7	0	5
MIGRANTE	Entero	1	99	99	9921	9	0	98
PAIS_NACIONALIDAD	Texto	Argentina	Venezuela	México	0	*NA*	*NA*	*NA*
PAIS_ORIGEN	Texto	Cuba	Venezuela	México	0	*NA*	*NA*	*NA*
UCI	Entero	1	99	97	2	97	0	91

Tabla 1: Resultado del análisis estadistico aplicado al conjunto de datos utilizado.

6.2. Visualización de datos

6.2.1. Gráfico de línea

La Figra 6 muestra el resultado de la implementación de un gráfico de línea en los paquetes de visualización de datos *Bokeh* (Figura 6a) y *Pygal* (Figura 6b).

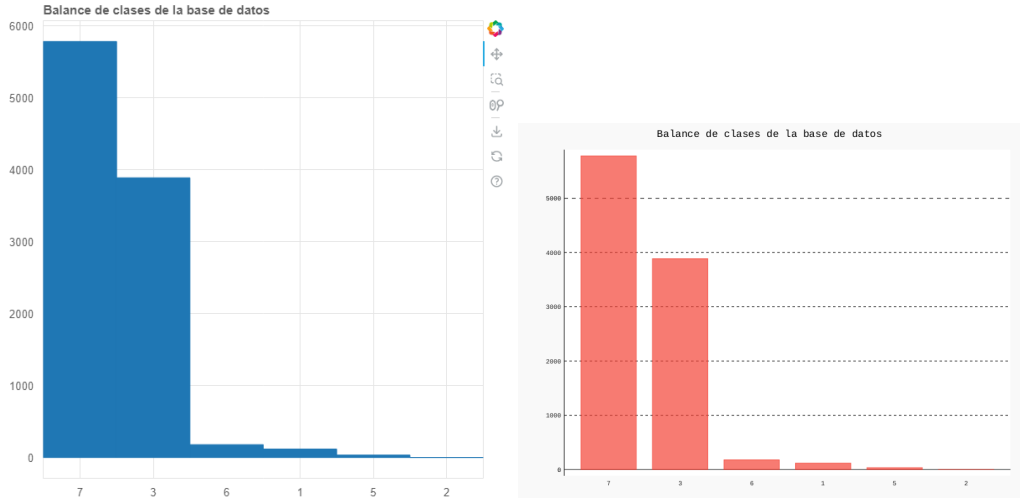


(a) Implementación en Bokeh (b) Implementación en Pygal

Figura 6: Implementación de gráfico de línea en diversos paquetes de visualización de datos.

6.2.2. Gráfico de barras

La Figra 7 muestra el resultado de la implementación de un gráfico de barras en los paquetes de visualización de datos *Bokeh* (Figura 7a) y *Pygal* (Figura 7b).

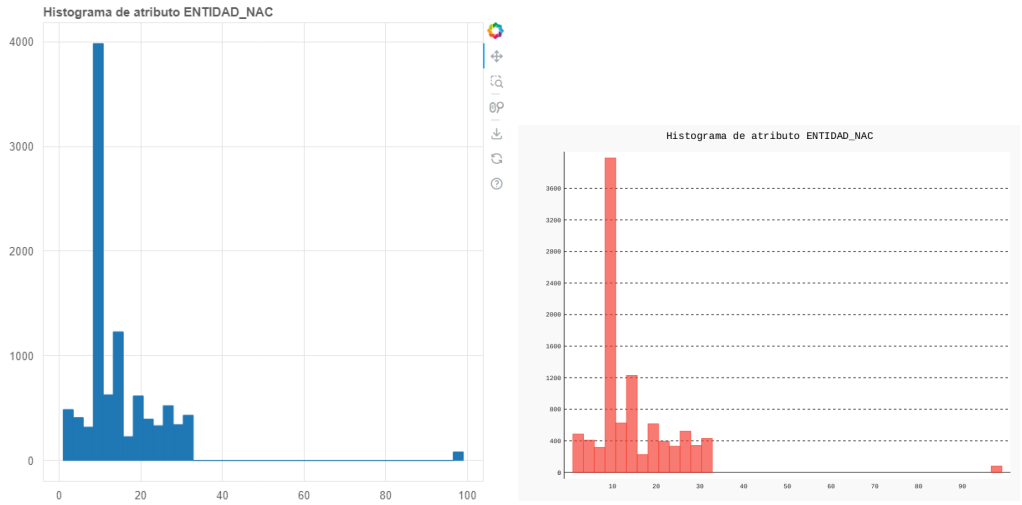


(a) Implementación en Bokeh (b) Implementación en Pygal

Figura 7: Implementación de gráfico de barras en diversos paquetes de visualización de datos.

6.2.3. Histograma

La Figra 8 muestra el resultado de la implementación de un histograma en los paquetes de visualización de datos *Bokeh* (Figura 8a) y *Pygal* (Figura 8b).

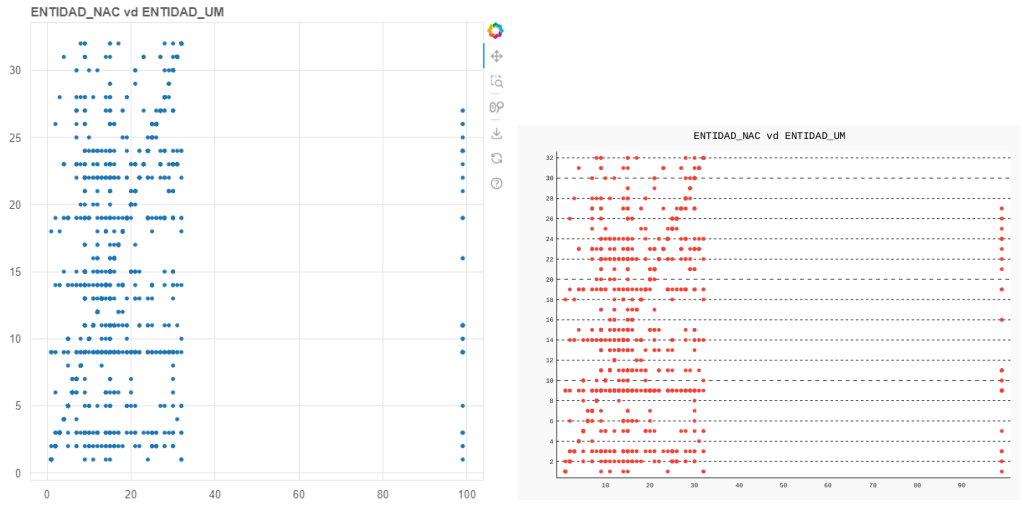


(a) Implementación en Bokeh (b) Implementación en Pygal

Figura 8: Implementación de histograma en diversos paquetes de visualización de datos.

6.2.4. Gráfico de dispersión

La Figra 9 muestra el resultado de la implementación de un gráfico de dispersión en los paquetes de visualización de datos *Bokeh* (Figura 9a) y *Pygal* (Figura 9b).

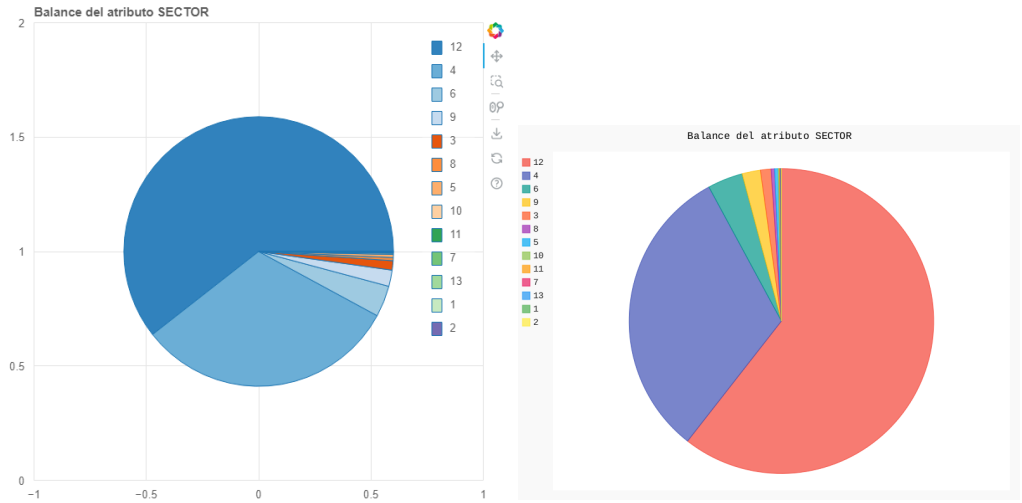


(a) Implementación en Bokeh (b) Implementación en Pygal

Figura 9: Implementación de gráfico de dispersión en diversos paquetes de visualización de datos.

6.2.5. Gráfico de pastel

La Figra 10 muestra el resultado de la implementación de un gráfico de pastel en los paquetes de visualización de datos *Bokeh* (Figura 10a) y *Pygal* (Figura 10b).

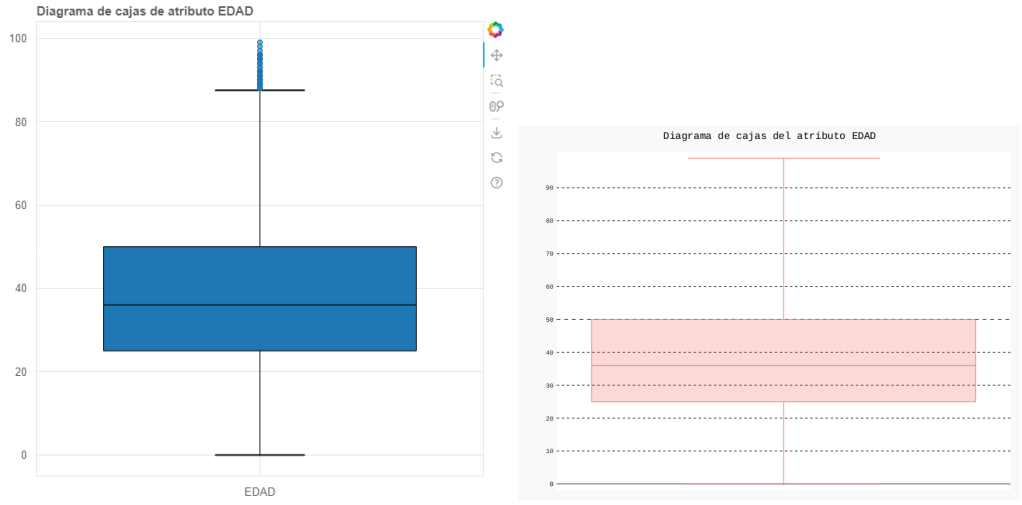


(a) Implementación en Bokeh (b) Implementación en Pygal

Figura 10: Implementación de gráfico de pastel en diversos paquetes de visualización de datos.

6.2.6. Diagrama de cajas

La Figra 11 muestra el resultado de la implementación de un diagrama de cajas en los paquetes de visualización de datos *Bokeh* (Figura 11a) y *Pygal* (Figura 11b).



(a) Implementación en Bokeh (b) Implementación en Pygal

Figura 11: Implementación de diagrama de cajas en diversos paquetes de visualización de datos.

7. Conclusiones

Es conocido que un modelo de inteligencia artificial es tan bueno como los datos con los que se haya entrenado, y para estar seguros de la calidad de los datos es necesario conocer su composición y medidas de estadistica descriptiva.

En esta práctica se logró obtener un panorama general del conjunto de datos utilizado mediante un análisis estadistico por atributo, obteniendo así valores máximos, mínimos, promedios, modas y medianas, desviaciones estándar y datos faltantes, entre otros.

Además, mediante la implementación de diversos gráficos fue posible obtener una representación visual de las relaciones existentes entre algunos atributos, así como la distribución de cada uno de los mismos.

Referencias bibliográficas

[1] M. A. Aceves Fernández, *Inteligencia artificial para programadores con prisa*. Universo de Letras, 2021, ISBN: 9788418854613.

[2] G. Westreicher, "Cuartil," Economipedia, Visitado: Febrero 2023. [En linea]. Disponible: https://economipedia.com/definiciones/cuartil.html

[3] J. M. Heras, "Correlación, covarianza e ibex-35," IArtificial.net, 2020, Visitado: Febrero 2023. [En linea]. Disponible: https://www.iartificial.net/correlacion-covarianza-ibex35

[4] ——, "Las 7 fases del proceso de machine learning," IArtificial.net, 2020, Visitado: Febrero 2023. [En linea]. Disponible: https://www.iartificial.net/fases-del-proceso-de-machine-learning

A. Código documentado

El código completo y funcional se puede encontrar anexo en el archivo *zip* compartido en conjunto con este reporte, así como también se puede encontrar dentro del repositorio de GitHub *MCI16-MachineLearning* dentro de la ruta *P1/Practica1_EnriqueMenaCamilo.ipynb*