Reporte Individual de Unidad I
Visualizando la función de densidad de probabilidad (PDF) y de distribución acumulada (CDF)

Torres Colorado Juan Daniel*
*Ingeniería en Tecnologías de la Información
Universidad Politécnica de Victoria

***Resumen*— Este proyecto subraya la importancia de las funciones de distribución acumulada (CDF) y las funciones de densidad de probabilidad (PDF) en el análisis de datos y la probabilidad. El objetivo principal fue desarrollar una visualización interactiva de las CDF y PDF utilizando el lenguaje de programación Python. Se implementaron clases y funciones para la manipulación y representación gráfica de distribuciones de probabilidad. La visualización no solo facilita la comprensión de estos conceptos fundamentales, sino que también ofrece una herramienta educativa valiosa para estudiantes y profesionales interesados en la estadística y el análisis de datos.**

I. Introducción

El documento del proyecto presente fue asignado por el docente de la asignatura "Sistemas Inteligentes" y tuvo lugar a mediados de mayo del presente año. El propósito de este proyecto tiene como motivo principal comprender la función de densidad de probabilidad (PDF) y la función de distribución acumulada (CDF) [1] através de una interfaz visual desarrollado en el lenguaje de programación Python3 [2] a partir del análisis de datos de un archivo *.ARFF* [3] proporcionado.

La **función de densidad de probabilidad** (PDF) es una función matemática que describe la probabilidad relativa de que una variable aleatoria continua tome un valor específico y representa cómo se distribuyen los valores de la variable a lo largo de sus posibles rangos. A pesar de no dar probabilidades exactas de valores específicos, resulta muy útil para la predicción de precipitaciones, el mercado de valores, la meteorología, etc.

La **función de distribución acumulada** (CDF) es una función matemática que describe la probabilidad de que una variable aleatoria tenga valores menores o iguales a un valor dado específico. Es decir, es una función acumulativa porque suma la probabilidad total hasta ese punto y su resultado siempre se encuentra entre 0 y 1.

Un **Formato de Archivo de Atributo-Relación** (ARFF) es un archivo de texto ASCII que describe una lista de instancias que comparten un conjunto de atributos y se utiliza para el preprocesamiento de datos, el intercambio de datos, aprendizaje automático, etc.

II. Desarrollo Experimental

El desarrollo experimental que se llevó en este proyecto se basa principalmente de las fuentes atribuidas por el docente, en tales casos, con el uso de la librería PyQt6 [4] para la creación de la interfaz interactiva, así como el uso de las librerías *numpy*, *pandas*, *scipy* y *matplotlib* [5, 6, 7, 8] para la visualización y análisis de datos. Además, gracias a los conocimientos previstos en antiguos proyectos similares, se obtuvo un cierto grado de facilidad para el desarrollo de la interfaz interactiva del usuario.

En primer lugar, para la comprensión de los conceptos de *PDF* y *CDF* no fue necesario el uso de investigación externa, esto debido a que en asignaturas pasadas se nos hizo de enseñanza estos conceptos matemáticos. Sin embargo, se llevó a cabo una investigación de un análisis acerca de la funcionalidad y uso de los archivos *ARFF*, esto con el propósito de obtener una comprensión sobre la estructura de datos bidimensional (*dataframe*) [9] con la que se pueda manipular para llevar a cabo el propósito de este proyecto.

En el desarrollo de este proyecto se hizo necesario la creación de diversas clases (tales son: *Window* y *ArffLector*) que contaran con la capacidad de gestionar la visualización en pantalla sobre la representación del conjunto de datos proporcionado. La figura 6 demuestra como se gestiona lo anterior dicho, en la cuál, el uso y la función de los diferentes métodos y variables serán explicados conforme los resultados se desarrollarán a continuación.

III. Resultados

Los resultados de este proyecto resultan en la visualización gráfica de la función de densidad de probabilidad (**PDF**) y la función de distribución acumulada (**CDF**) apartir del **dataframe** de un archivo **ARFF** proporcionado. En la figura 1 demuestra la ventana inicial contando con los elementos:

- **Botón** "Cargar Archivo": Este botón tiene la funcionalidad de demostrar una ventana emergente con la capacidad de seleccionar un sólo archivo de formato *.arff*.
- **ComboBox**: Una vez proporcionado el archivo requerido, este tiene la funcionalidad de seleccionar a cuál de las columnas numéricas se quiere visualizar la figura correspondiente.

En la figura 2 se representa la ventana en la que, gracias a la clase *QFileDialog* de la librería de *PyQt6*, permite que la selección del archivo *.arff* se proporcione de una manera más fácil y sencilla de obtener.

Una vez que el archivo ha sido proporcionado correctamente, con el uso de la clase *arff* de *scipy.io* para cargar el archivo,

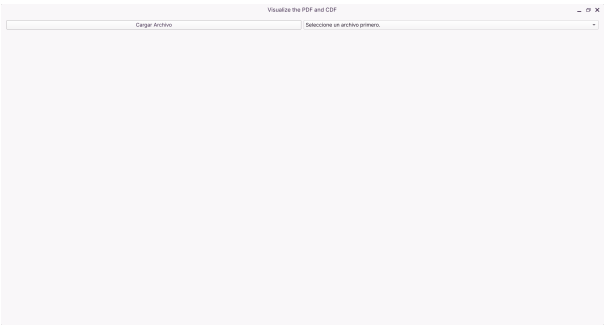


Figura 1: Ventana Inicial

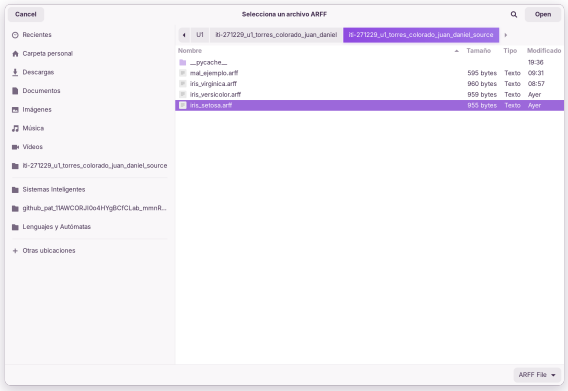


Figura 2: Selección de Archivo

la clase *DataFrame* de *pandas* para obtener la estructura de datos bidimensional, las funciones de *numpy* para obtener las columnas numéricas del dataframe, crear el histograma [10] y la *CDF*, la clase *FigureCanvasQTAgg* y *Figure* de *matplotlib* para la creación de la representación, se visualiza en la ventana principal la figura obtenida a partir de distintos procesos de validaciones este conjunto de datos. El resultado de lo anterior mencionado se aprecia en la figura 3.

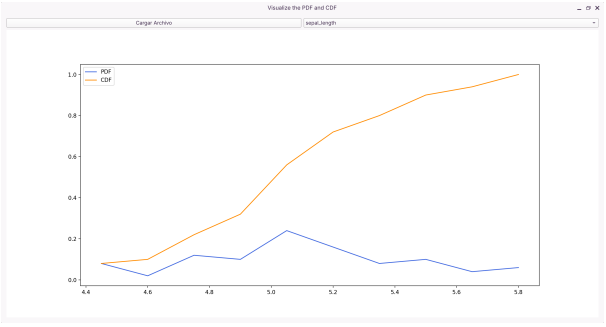
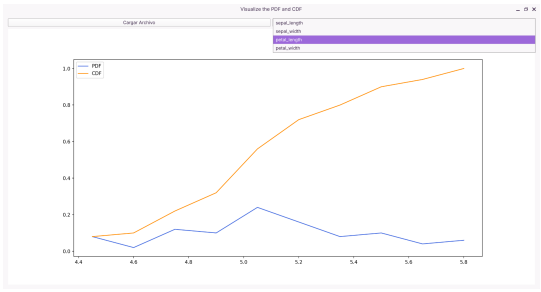


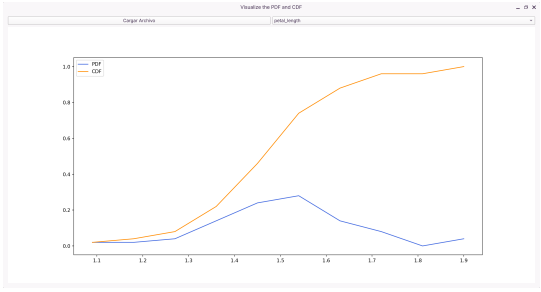
Figura 3: Archivo y Figura cargada

Si en el archivo proporcionado se cuenta con una o menos de diez columnas numéricas, entonces se puede seleccionar cuál de estas se quiere hacer la visualización de sus conjuntos de datos. En la figura 4 se demuestra este proceso.

En dado caso de que se decida cargar un nuevo archivo



(a) Selección de columna



(b) Representación de la selección de columna

Figura 4: Cambio de figura

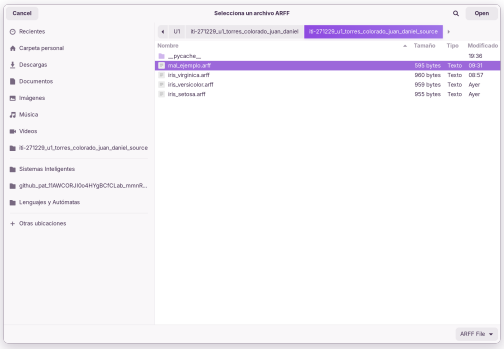
pero este resulte ser inválido o no fue seleccionado, entonces se mostrará un mensaje de advertencia indicando dicho error y la figura será eliminada en el caso de haber una ya previamente cargada. Esto se puede apreciar en la figura 5.

IV. Conclusión

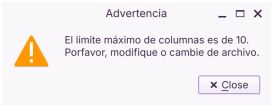
En este trabajo se desarrolló una interfaz visual en Python3 utilizando la librería PyQt6 para la visualización de funciones de densidad de probabilidad (PDF) y de distribución acumulada (CDF) a partir de datos proporcionados en archivos de formato ARFF. El proyecto tuvo como objetivo principal la comprensión y representación gráfica de estas funciones, facilitando el análisis de datos a través de una herramienta interactiva.

Se hicieron uso de diversas librerías de Python como numpy, pandas, scipy y matplotlib para la manipulación y visualización de datos. Se diseñaron y programaron clases específicas como *Window* y *ArffLector* para gestionar la interfaz y el procesamiento de los datos mediante la carga de un archivo ARFF y la selección dinámica de columnas numéricas mostrando las gráficas correspondientes y presentarlos de una manera accesible y comprensible para cumplir con el objetivo del proyecto. Además, se implementaron funcionalidades para manejar archivos inválidos y proporcionar mensajes de advertencia al usuario cuando sean necesarias.

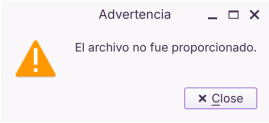
Para finalizar, este proyecto proporcionó conocimiento de manera exitosa al alumno, además de proporcionar resultados satisfactorios para la visualización de las funciones matemáticas PDF y CDF.



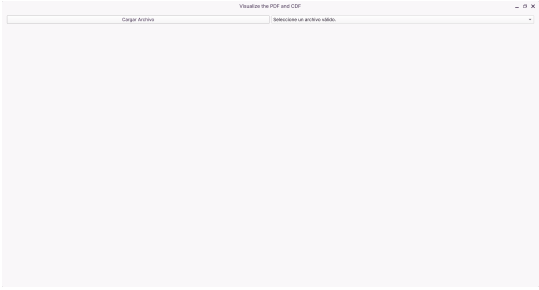
(a) Selección de un archivo inválido



(b) Advertencia de archivo inválido



(c) Advertencia de archivo no seleccionado



(d) Figura no cargada

Figura 5: Selección errónea o De-selección de Archivo

Referencias

[1] Kristin Potter et al. "Interactive Visualization of Probability and Cumuluative Density Functions". En: *International Journal for Uncertainty Quantification* 2.4 (2012), págs. 397-412. ISSN: 2152-5080.

[2] Guido Van Rossum y Fred L. Drake. *Python 3 Reference Manual*. Scotts Valley, CA: CreateSpace, 2009. ISBN: 1441412697.

[3] Weka. *Attribute-Relation File Format (ARFF)*. https://ml.cms.waikato.ac.nz/weka/arff.html. Consultado el 04-06-2024.

[4] Riverbank Computing Limited. *PyQt6 - Comprehensive Python Bindings for Qt v6*. https://pypi.org/project/PyQt6/. Consultado el 04-06-2024.

[5] Charles R. Harris et al. "Array programming with NumPy". En: *Nature* 585.7825 (sep. de 2020), págs. 357-362.

[6] Inc. pandas via NumFOCUS. *pandas - Python Data Analysis Library*. https://pandas.pydata.org/. Consultado el 04-06-2024.

[7] Pauli Virtanen et al. "SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python". En: *Nature Methods* 17 (2020), págs. 261-272.

[8] J. D. Hunter. "Matplotlib: A 2D graphics environment". En: *Computing in Science & Engineering* 9.3 (2007), págs. 90-95.

[9] DataScientest. *¿Qué es un DataFrame?* https://datascientest.com/es/que-es-un-dataframe. Consultado el 04-06-2024.

[10] GCFGlobal. *Histograma de datos*. https://edu.gcfglobal.org/es/estadistica-basica/histograma-de-datos-/1/. Consultado el 04-06-2024.

V. ANEXOS

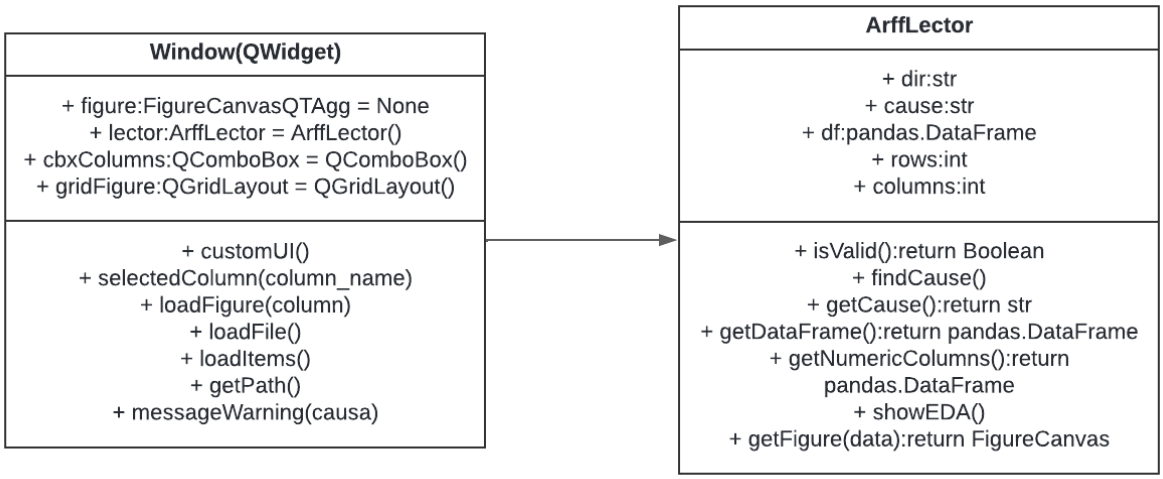


Figura 6: Diagrama de clases