Uso de Kaggle en la Enseñanza de Estadística Descriptiva: Una Experiencia de Clase

Gladys Denisse Salgado Suárez [1], José Rubén Conde Sánchez[2], Guillermina Sánchez López[3]
[1]Centro de Investigación y de Estudios Avanzados del IPN, Ciudad de México, México. gladys.salgado@udlap.mx
[2] Facultad de Ciencias Físico Matemáticas, Benemérita Universidad Autónoma de Puebla, Puebla, México. rconde@fcfm.buap.mx
[3]Preparatoria Regional "Simón Bolivar", Benemérita Universidad Autónoma de Puebla, Atlixco, Puebla, México. guillermina.sanchez@correo.buap.mx

Resumen

Los modelos didácticos tradicionales cada día se transforman con las nuevas corrientes educativas, es recurrente en los nuevos modelos considerar la participación del alumno como eje central de la enseñanza-aprendizaje. La permanente búsqueda de metodologías de enseñanza determina que los alumnos son los actores principales de un sistema educativo que transforma el tradicional paradigma donde el docente es el centro del conocimiento. De esto, las experiencias de clase son una forma de exponer los logros de los alumnos a través de actividades en que ellos sean el eje central y no el docente. En este trabajo, se expone una experiencia de clase en la enseñanza de la estadística descriptiva, a través del uso de *datasets* de *Kaggle*.

Palabras clave: Estadística, kaggle, tecnología.

Abstract

The traditional didactic models are transformed every day with the new educational currents, it is recurrent that the new models consider the participation of the student as the central axis of teaching-learning. The permanent search for teaching methodologies determines that students are the main actors in an educational system that transforms the traditional paradigm where the teacher was the center of knowledge. From this, class experiences are a way of exposing the achievements of students through activities in which they develop. In this work, class experience in the teaching of descriptive statistics is exposed, using Kaggle datasets.

Keywords: Statistics, kaggle, technology.

Modalidad: Ponencia.

I. Introducción

Parte importante del quehacer docente está enfocado en la búsqueda del logro de los aprendizajes esperados en los estudiantes. Por otro lado, el uso de las Tecnologías de la Información y Comunicación (TICs), se ha convertido en un aliado en la práctica de enseñanza. Las herramientas tecnológicas representan una oportunidad para complementar los temas teórico-prácticos en el aula [4], y ahora en el tiempo que estamos viviendo, también a distancia o modalidad híbrida. El uso en clase de dispositivos electrónicos por parte del docente y del estudiante, es cada vez una necesidad que requiere de una conectividad estable al internet y de buena tasa de transferencia de datos para el acceso a plataformas educativas alojadas en el internet o en la nube.

La tecnología es un referente para usarse cada vez más como herramienta en clase, en el caso del alumno, se aprovecha el momento tecnológico en el que está inmerso, por ende, se valoran y aprovechan las habilidades tecnológicas adquiridas; y, para el docente, es una oportunidad para explorar en la diversidad de metodologías existentes. A la par, la tecnología educativa reconoce el uso de las TICs en el aula, para acercarse poco a poco a casi todas las materias curriculares en particular en las matemáticas promoviendo mejores emociones y actitudes hacia ellas, además de ayudar a desarrollar el razonamiento matemático y habilidades del pensamiento [5]. Con esto proponemos y organizamos una experiencia de clase para la enseñanza de la estadística descriptiva que cree una necesidad al alumno de usar y aplicar conceptos que son parte de sus conocimientos previos.

La experiencia de clase presentada en este trabajo busca generar un ambiente en donde el docente incentive la curiosidad o necesidad por parte del alumno a través de preguntas de conocimiento general sobre sus intereses personales, como: deportes, juegos, ciencia, películas, redes sociales, entre otros temas. Los datos o conjunto de datos, *datasets*, son descargados de la empresa *Kaggle*[1]. En *Kaggle* "encontrará los datos que necesita para hacer su trabajo de ciencia de datos. Utilice más de 50 000 conjuntos de datos públicos y 400 000 cuadernos públicos para conquistar cualquier análisis en poco tiempo." Los datos versan en categorías como: Ciencia Computacional, Educación, Clasificación, Visión Computacional, Procesamiento de Lenguaje Natural, Visualización de Datos, Modelo de Datos Pre Entrenados, Pronósticos, etc. Es una fuente vasta de *datasets* para analizar, se podría considerar que existen categorías para cualquier área de interés, lo que permite a cada alumno analizar los datos que son realmente de su interés. El termino *dataset* aparece con la llegada de las emergentes tecnologías como el Big Data, son conjuntos de datos estructurados accesibles desde cualquier lenguaje de programación (diccionarios de Python, formatos *json* , etc.), se refiere a una única base de datos autorelacionada o relacionada con otras bases de datos, cada columna del *dataset* representa una variable y cada fila corresponde a cualquier dato. Su utilidad se ha vuelto importante ya que diversos centros de investigación científica, instituciones financieras, centros médicos y muchos más ofrecen *dataset* públicos o privados con la finalidad de que cualquier investigador o estudiante los utilice para probar algoritmos, practiquen con ellos sin la necesidad de preocuparse por la calidad de los datos[2], etc.

[1] https://www.kaggle.com/

[2] A Beginner's Guide to Sentiment Analysis with Python, 2020)

II. Fundamento de la experiencia de clase

El objetivo de la experiencia de clase está basado en emplear los estándares ISTE[3] (Sociedad Internacional para la Tecnología en la Educación), se busca que el docente ayude a los estudiantes e impulsar su propio aprendizaje con el uso de la tecnología, así, lograr justificar el uso y experiencia del alumnado; además, aprovechar la disposición de los datos libres en el internet y utilizarlos en la enseñanza de los conceptos básicos de estadística descriptiva en nivel de licenciatura.

En esta experiencia de clase, se ve reflejado, el proceso de metacognición, que en la literatura de investigación educativa otorga definiciones diversas sobre los procesos (conocimiento, regulación y experiencia), que giran en torno a lo siguiente: "capacidad de una persona de pensar en su pensamiento de conocer sus conocimientos, de ser consciente de su conciencia"[4]. La metacognición está apoyada con las metodologías de enseñanza orientadas a fortalecer los aprendizajes esperados planteados. Un punto importante que se resalta es el proceso reflexivo sobre lo que ha logrado el estudiante al utilizar los *dataset* en el estudio que requieren en esta experiencia de clase, y se orienta hacia el interés en abarcar otro tipo de *dataset*, ya que se vale de las herramientas que le ofrece la estadística descriptiva para lograr analizar otras áreas de la ciencia.

III. Diseño e implementación de la planeación

Para este estudio, se considera y selecciona el programa de la materia Adquisición y Procesamiento de Datos Experimentales (APDE) correspondiente al plan de la licenciatura en Física de la Facultad de Ciencias Físico Matemáticas (FCFM) de la Benemérita Universidad Autónoma de Puebla (BUAP). Para mostrar la experiencia de clase, se seleccionó este programa, ADPE, porque el plan de la materia lo permite al contener el tema de análisis de datos. Esto da la oportunidad de enriquecerla la materia con herramientas de Tecnología Educativa; abordando así, estrategias de enseñanza aprendizaje que tienen detrás toda una base pedagógica conceptual en cada uno de los momentos educativos y favorecer la práctica docente. Se espera fortalecer los aprendizajes por parte del alumnado. Es importante considerar que la materia seleccionada corresponde a séptimo semestre de la carrera, los alumnos han aprobado la materia de Física Computacional y han tenido experiencia en programación básica en lenguaje de programación Python, lo que infiere que conocen de cierto modo un lenguaje de programación. El desarrollo de esta experiencia de clase es desarrollado en el tiempo de retorno hibrido a la universidad toda vez que los alumnos han sido vacunados contra el COVID-19, esto quiere decir que el curso se lleva acabó en forma síncrona y asíncrona. El alumnado contaba con las grabaciones de las reuniones.

La Tabla 1 describe la planeación de las actividades, que en sí es la metodología propuesta para desarrollar la experiencia de clase. En la sección de objetivos, se muestra una parte de importancia para este trabajo, la conjunción entre el tema central de la materia APDE con la asociación de programación en un lenguaje de programación sugerido como Python, lo que permite emplear recursos como implementación de funciones, ciclos, procesamiento y visualización de la

[3] https://www.iste.org/es/
[4] https://www.ceupe.com/blog/que-son-los-procesos-metacognitivos.html

información. Python les ofrece la característica de poder realizar programación que es considerada como computo en la nube, si ésta es realizada desde *Google Colabatory*[5].

Tabla 1 Planeación de la experiencia de clase

Experiencia de Clase	Uso de *dataset* para la enseñanza de Estadística Descriptiva	
Objetivo	Aprovechar la experiencia acumulada de los estudiantes en el uso de las herramientas tecnológicas como la computadora, laptop o teléfonos inteligentes y la accesibilidad a la información dispuesta en el internet. Utilizar datasets de diversas áreas con la idea de captar un mayor interés en los alumnos. Entrenarse con los datos en el aprendizaje de los conceptos básicos de estadística descriptiva Utilizar los recursos de la estadística descriptiva para procesar, analizar *dataset* ofrecidos *Kaggle* para estimular el aprendizaje significativo en el alumnado. Aprender la secuencia básica para realizar computo en la nube.	
Temas transversales	Estadística descriptiva Física Computacional, Programación Python – estructura de datos; además, del área que corresponde a la categoría del *dataset* que elige el alumno (Medicina, Sociología, Arte, etc.)	
Estándares ISTE	**2.6 Facilitador** Los educadores facilitan el aprendizaje con el uso de la tecnología para apoyar el logro académico de los estudiantes. Para esto, los educadores: Crean oportunidades de aprendizaje que desafían a los estudiantes a utilizar un proceso de diseño y de pensamiento computacional para innovar y solucionar problemas. **2.6.b** Gestionan el uso de la tecnología y las estrategias de aprendizaje de los estudiantes en plataformas digitales, entornos virtuales, espacios de creación prácticos o en el campo.	
Metodología	Trabajo colaborativo / Aprendizaje Basado en Problemas (ABP) (La experiencia de clase forma parte del proyecto de clase, por esto se considera la metodología el ABP).	
Nivel de inserción de la tecnología	_____ Sustitución _____ Argumento ___X_ Modificación _____ Redefinición	
Secuencia didáctica		**Materiales**
Sesión 1-2 (2hrs/sesión) Que el estudiante: 1. Conozca la plataforma de *Kaggle* 2. Identificar categoría del *dataset* de su interés 3. Repace las estructuras básicas de Python para manejo de diccionarios y formato de datos tipo *json*. 4. Explore las características del *dataset* en cuanto a número de datos, curar datos. 5. Acondicione el *dataset* en el drive de Google toda vez que el proyecto se realiza, de preferencia, en *Google Colab* **Sesión 3-4 (2hrs/sesión)** Que el Estudiante: 1. Una vez que haya agregado el *dataset* en el drive, abra el archivo en *Google Colab*. 2. Que conozca el archivo, haciendo uso de operaciones básicas como: cuántas características tiene el archivo (columnas), salidas (*outputs*), tamaño de características, tipo de datos, etc. 3. Segmentar el archivo, graficar grupos de características. 4. Implementar las funciones de la estadística descriptiva: suma, promedio, correlaciones, desviación standard, etc. 5. Comparar las funciones implementadas con las funciones desarrolladas con módulos como: *numpy*, *scipy* u otro 6. Construya gráficas de los resultados que vaya generando. **Sesión 5 (2hrs/sesión)** Que el estudiante: 1. Comparta con otros compañeros sus códigos e ilustraciones para recibir retroalimentación. 2. Una vez que ha realizado las correcciones necesarias, comience a realizar el reporte de su proyecto entregable. 3. Presente su trabajo con el resto del grupo. 4. Cargue su proyecto en Tareas de Teams.		IDE de Python Computadora Conectividad a Internet

[5] https://colab.research.google.com/?utm_source=scs-index

Evaluación	Producción gráfica, escrita y oral.	**Transversalidad**	Tecnología Ciencias Ingeniería

El desarrollo de la experiencia de clase es mostrado en la Tabla 2, en ella se muestran fases como planteamiento, seguimiento, evaluación, retroalimentación, entre otros.

Tabla 2 Implementación y Evaluación de experiencia de clase con trabajo síncrono y asíncrono para el aprendizaje.

Campo de formación académica	Pensamiento matemático-computacional a través de la manipulación y procesamiento de *dataset* en la enseñanza de la estadística descriptica	**Semana**	5 sesiones
Aprendizajes esperados	1. Reconocer y emplear los conceptos y fórmulas básicas de la estadística descriptiva. 2. Proponer una acción que conduzca a alguna mejora en los datos basado en los resultados arrogados.	**Competencias**	Conocer los conceptos básicos de la estadística descriptica como población, muestra, promedio, desviación estándar moda, estos datos, son bien representados en tablas de frecuencias y mostrados en diversos gráficos que muestran adecuadamente la tipología de datos procesados.
Ejes	2. Desarrollo de Habilidades del Pensamiento Complejo (DHPC) 3. Desarrollo de Habilidades en el uso de la Tecnología, la Información y la Comunicación (DHTIC) 4. Lenguas Extranjeras 5. Educación para la Investigación	**Tema:**	Estadística Descriptiva
Estándar ISTE	Los educadores facilitan el aprendizaje con el uso de la tecnología para apoyar el logro académico de los estudiantes, mediante la puesta en práctica de los Estándares ISTE • 2.6.b Gestionan el uso de la tecnología y las estrategias de aprendizaje de los estudiantes en plataformas digitales, entornos virtuales, espacios de creación prácticos o en el campo.	**Metodología Educativa**	• ABP
Sesiones distribuidas			
Clase sincrónica		**Clase asincrónica**	
Que el estudiante: 1. Revise y repase los materiales de programación en lenguaje Python para agilizar el uso de sus recursos a la hora de implementar los algoritmos para calcular las operaciones involucradas. 2. Una vez que el profesor haya expuesto algún caso de uso en que el alumno se pueda guiar en el transcurso de la actividad por realizarse. 3. El alumno indaga en la página web de *Kaggle*, elije el *dataset* de su interés. 4. Se descarga y aloja el *dataset* en su sesión de la nube en el entorno de *Google Colab*, o en el entorno de desarrollo que prefiera. 5. Realizar las operaciones de la estadística descriptica con los datos seleccionados.		Que el estudiante: 1. Vea nuevamente el video del caso de uso expuesto por el docente y rehaga el proyecto empleando su *dataset*. 2. Asociarse con algún compañero de clase, aclarar dudas, además de fomentar la colaboración entre compañeros. 3. Redactar una experiencia sobre el uso de los recursos a través de los *dataset*.	

6. Exponer sus resultados .			
Evaluación	**Recursos y materiales**	**Herramientas tecnológicas**	**Adecuaciones curriculares**
Comprobar los resultados arrogados de la Implementación de los programas en Python con programas basados en el uso de funciones, como MATLAB, R o el mismo Python con módulos como: *scipy*, *numpy*, etc.	• Laptop, celular, tablet • Sesión en *Google colab* • Código cargado en la nube • Libro de texto	• Microsoft Teams • Libreta digital de One note • *Google Colab*	• El profesor hace el planteamiento de la actividad a través de TEAMs, y muestra un ejemplo de uso del *dataset* desde *Kaggle*.
Nivel de inserción de la tecnología	____ Sustitución ____ Argumento ____ Modificación _X__ Redefinición		
Consideraciones para el día siguiente Repasar los conceptos y las ecuaciones de la estadística descriptiva			

Al alumno se le muestra un *Caso de Uso* que le sirve como referencia, se expone una situación que sirva para guiarse en una metodología ordenada en una serie de pasos. De esta forma el alumno tiene un ejemplo que le sirve como referencia; una vez comprendido por parte del alumno, hará lo necesario para seguir lo requerido en la Tabla 2.

IV. Resultados

Se muestra solo un caso de varios presentados en clase por parte de los alumnos, el cual ilustra los resultados al emplear los conceptos de estadística descriptiva bajo la experiencia de clase aquí presentada.

De la base de datos *Pima Indians Diabetes*[6] del *National Institute of Diabetes and Digestive and Kidney Diseases*, "el objetivo del *dataset* es predecir de manera diagnóstica si un paciente tiene o no diabetes, en función de ciertas medidas de diagnóstico incluidas en el conjunto de datos. Se impusieron varias restricciones a la selección de estas instancias de una base de datos más grande. En particular, todos los pacientes aquí son mujeres de al menos 21 años de herencia indígena pima"[6]. Para este caso, se involucra al estudiante en una aplicación directa de la descripción de los datos usando la estadística directamente sobre ellos.

Los datos están en términos de nueve características:

I. *Embarazos*: número de veces embarazadas.
II. *Glucosa*: concentración de glucosa en plasma a las 2 horas en una prueba de tolerancia oral a la glucosa.
III. *Presión arterial*: presión arterial diastólica (mm Hg).
IV. *Grosor de la piel*: Grosor del pliegue cutáneo del tríceps (mm).
V. *Insulina*: insulina sérica de 2 horas (mu U/ml).
VI. *IMC*: Índice de masa corporal (peso en kg/(altura en m)^2).
VII. *DiabetesPedigreeFunction*: función de pedigrí de diabetes.
VIII. *Edad*: Edad (años).
IX. *Resultado*: variable de clase (0 o 1).

La Ilustración 1 muestra los primeros datos del archivo, las operaciones que se realizan a continuación están referenciadas a cuestiones más técnicas como la operatividad del lenguaje de programación que permite conocer elementos que son parte del grupo de datos como: tamaño de datos, tipo de datos, variables, dimensión de la tabla, tamaño de dato.

[6] https://data.world/data-society/pima-indians-diabetes-database

Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFu	Age	Outcome
6	148	72	35	0	33.6	0.627	50	1
1	85	66	29	0	26.6	0.351	31	0
8	183	64	0	0	23.3	0.672	32	1
1	89	66	23	94	28.1	0.167	21	0
0	137	40	35	168	43.1	2.288	33	1

Ilustración 1 Primeros datos del dataset diabetes.

De la tabla de datos se hace estadística descriptiva, calculando promedio, moda, desviación stantard, o incluso ajuste de datos e interpretando los resultados de acuerdo al tema. Por otro lado, realizar estas operaciones le permite o le exige también al alumno, utilizar operaciones de selección de rangos sobre una matriz para segmentar y procesar los datos. Por ejemplo: graficar sólo una característica como edad en función de la glucosa, etc.

Existen herramientas que permiten hacer un resumen analítico de los datos, con herramientas como *pandas_profiling*[7], que, si bien este módulo está basado en *pandas*, utilizarlo representa una considerable ayuda para cotejar los resultados de las ecuaciones con los valores mostrados por *pandas_profiling*.

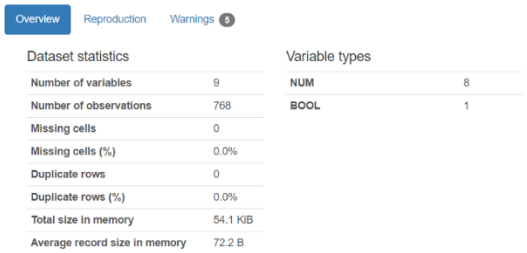

Ilustración 2 import pandas_profiling as pp, pp.ProfileReport(df)

El recurso de *pandas_profiling* es de enorme ayuda a la hora de cotejar los resultados de las funciones creadas con Python con los datos mostrados en el reporte *profiling*, de igual manera permite hacer el perfil a través de un reporte de cada una de las variables.

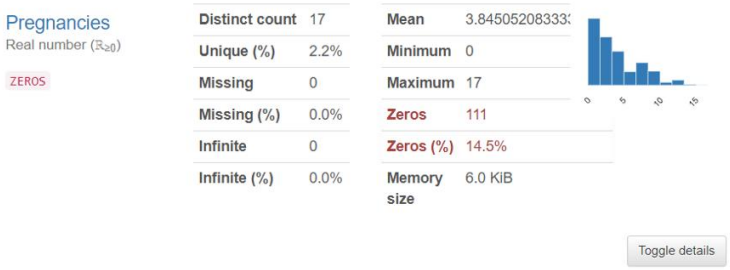

Ilustración 3 Perfil de la característica Pregnancies.

[7] https://pandas-profiling.ydata.ai/docs/master/index.html

De esta forma se realiza la exploración de cada una de las variables, permitiendo posteriormente graficar varias variables en un mismo gráfico, y explorar cuestiones como la correlación entre variables.

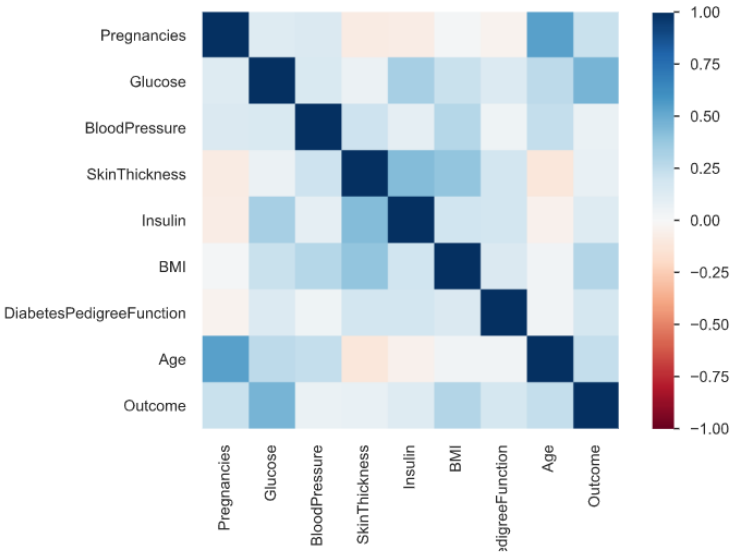

Ilustración 4 Correlación entre variables.

V. Conclusiones

Cada alumno presenta *dataset* diferentes, lo que denota que sus intereses son diferentes, esto fortalece los conceptos teóricos expuestos por el docente a través de la atención natural de estudiante en analizar datos de su propia elección.

Revisar el tema de la estadística descriptiva a través de Tecnología Educativa representa un reto para el docente, ya que logra el dominio no sólo de los conceptos sino también de la tecnología.

Se utilizan diversas herramientas para favorecer el pensamiento crítico computacional que permitan expresar ideas y experiencia en el proceso de aprendizaje.

VI. Bibliografía

[1] Todd L. Pittinsky, Science, Technology, and Society: New Perspectives and Directions; Cambridge Handbooks in Psychology, Cambridge University Press. 2019.

[2] García, F. “Los modelos didácticos como instrumento de análisis y de intervención en la realidad educativa”. En: Revista Bibliográfica de Geografía y Ciencias Sociales. Universidad de Barcelona . Nº 207, 18 de febrero 2000.

[3] Homero, G., Sosa, M.R., y Martínez, F. "Modelos didácticos en la educación superior: una realidad que se puede cambiar". Revista de Currículum y Formación de Profesorado, 22(2), 447-469, 2018.

[4] UNESCO (2021). (20 de septiembre, 2021) *Las TIC en la Educación*. UNESCO. https://es.unesco.org/themes/tic-educacion

[5] Orozco, C. y Labrador, M. E. (2006). La tecnología digital en educación: Implicaciones en el desarrollo del pensamiento matemático del estudiante. *Theoria.* 15(2), 81-89.