TRABAJO DE GRADO
Opción Seminario-Diplomado.

ANÁLISIS COMPUTACIONAL DE BASE DE DATOS ESTUDIANTIL, UTILIZANDO ALGORITMOS DE MACHINE LEARNING

Corporación Universitaria Remington.

Nombre de la facultad: Ingenierías

Nombre del programa académico: Ingenierías de sistemas

Estudiantes:

Franklin Yesid Artunduaga Mendoza.

Elkin Pulgarín Serna.

Julián Andrés Giraldo Vásquez.2

Tutor: Juan Carlos Briñez de León

Opción de Trabajo de grado Seminario-Diplomado.

2024.

Dedicatoria

Con esfuerzo y dedicación hemos llegado al momento más esperado de nuestras carreras muchas gracias a nuestras familias por el apoyo directo e indirecto que nos prestaron, lo logramos gracias, Julián.!! Elkin.!! Franklin.!!.

Agradecimientos

Gracias a los docentes de la antigua sede Cartago por ese empeño y amabilidad que nos brindaron en estos años con sus enseñanzas, sus consejos, que gracias a eso hoy podemos decir que estamos a un paso, gracias a los docentes de sede Pereira que nos recibieron con los brazos abiertos para continuar nuestro proceso académico, no sin antes un reconocimiento al docente **Juan Carlos Briñez de León** por compartir su amplia experiencia académica que nos permitió abrir los ojos al mercado actual, nos hubiera encantado tenerlo más tiempo en nuestros proceso académico, gracias profe…!!.

Tabla de Contenidos

Resumen

El trabajo, está basado en analizar la base de datos de los estudiantes de Bangladesh, en 5 áreas específicas (Matemáticas, Ciencias, Ciencias Sociales, inglés y Arte y cultura) en estos podemos encontrar los nombres de cada estudiante los cuales son jóvenes que cursan educación media, y están ad-portas de salir de la institución educativa. Esta base de datos nos muestra los puntajes comparativos entre mujeres y hombres, con los cuales, podemos deducir el nivel académico de cada estudiante.

En este trabajo, se ejecutan unos algoritmos que nos permiten modificar los datos de nuestra base estudiantil, para poder realizar el tratamiento de los datos, desarrollando algoritmos tales como “eliminar una columna” o “cambiar datos categóricos a numéricos” o “verificar información de la tabla”, para posteriormente ejecutar algoritmos que nos muestran estadísticas que representan la información más relevante de nuestra base de datos, tales como, matrices de correlación y gráficas de dispersión, de columnas o en torta, para exponer dichas estadísticas.

Luego se procede a realizar algoritmos de **modelos de regresión** en los que se puede observar los márgenes de error dentro de la base de datos, como el modelo KNN y el modelo ANN, en donde se evalúa una métrica de 3 datos que se quieran consultar fuera de los ya relacionados en la base de datos para obtener un promedio deseado con los datos escritos para predecir un promedio final.

Nuestra recomendación en este trabajo es analizar el rendimiento académico de cada estudiante para visualizar las posibles falencias o virtudes, para obtener estadísticas que nos permitan identificarlas y así saber que mejoras se pueden efectuar en el programa académico.

Palabras clave

Tratamiento de los datos, datos categóricos, base de datos, matrices de correlación, gráficas, KNN y ANN, estadísticas, modelos de regresión, márgenes de error

1. Marco conceptual y contextual

1.1 Contexto:

1.1.1 Sistemas de recomendación.

Análisis de Datos Educativos: El análisis de datos en el ámbito educativo, conocido como "learning analytics", implica la recopilación y análisis de datos sobre estudiantes para mejorar los procesos de enseñanza y aprendizaje. Este enfoque permite identificar tendencias, predecir resultados académicos y personalizar la educación según las necesidades individuales. (EducaOpen, s.f.)

El Machine Learning es una disciplina del campo de la Inteligencia Artificial que, a través de algoritmos, dota a los ordenadores de la capacidad de identificar patrones en datos masivos y elaborar predicciones (análisis predictivo). Este aprendizaje permite a los computadores realizar tareas específicas de forma autónoma, es decir, sin necesidad de ser programados. (iberdrola)

El término se utilizó por primera vez en 1959. Sin embargo, ha ganado relevancia en los últimos años debido al aumento de la capacidad de computación y al boom de los datos. Las técnicas de aprendizaje automático son, de hecho, una parte fundamental del Big Data. (iberdrola)

El algoritmo k vecinos más cercanos (KNN) es un clasificador de aprendizaje supervisado no paramétrico, que emplea la proximidad para realizar clasificaciones o predicciones sobre la agrupación de un punto de datos individual. Es uno de los clasificadores de clasificación y regresión más populares y sencillos que se emplean actualmente en el machine learning. (Ellucian)

El algoritmo de k vecinos más cercanos, también conocido como KNN o k-NN, es un clasificador de aprendizaje supervisado no paramétrico, que utiliza la proximidad para hacer clasificaciones o predicciones sobre la agrupación de un punto de datos individual. (IBM)

Si bien se puede usar para problemas de regresión o clasificación, generalmente se usa como un algoritmo de clasificación, partiendo de la suposición de que se pueden encontrar puntos similares cerca uno del otro. (IBM)

Funcionamiento de ANN

Vectorización: cada punto de datos del conjunto de datos se representa como un vector en un espacio multidimensional. (learnmicrosoft)

Indexación y estructuras de datos: los algoritmos ANN usan estructuras de datos avanzadas (por ejemplo, árboles KD, hash con distinción de localidad o métodos basados en grafos) para indexar los puntos de datos, lo que permite búsquedas más rápidas. (learnmicrosoft)

Cálculo de distancia: en lugar de calcular la distancia exacta a cada punto, los algoritmos ANN usan la heurística para identificar rápidamente las regiones del espacio que probablemente contienen los vecinos más cercanos. (learnmicrosoft)

Búsqueda de vecinos: el algoritmo identifica un conjunto de puntos de datos que probablemente están cerca del punto de consulta. No se garantiza que estos vecinos sean los puntos más cercanos exactos, pero son lo suficientemente cercanos para fines prácticos. (learnmicrosoft)

Realizar predicciones:

Clasificación: para las tareas de clasificación, ANN asigna la etiqueta de clase al punto de consulta más común entre los vecinos identificados, de manera similar a kNN.

Regresión: para las tareas de regresión, ANN predice el valor del punto de consulta como el promedio (o la media ponderada) de los valores de los vecinos identificados. (learnmicrosoft)

1.1.2 Algoritmos de Machine learning en sistemas de recomendación.

Aplicación de Machine Learning en la Educación: La integración de técnicas de machine learning en el análisis de datos educativos facilita la identificación de patrones complejos y la predicción de resultados académicos. Por ejemplo, algoritmos de clasificación pueden predecir el rendimiento de los estudiantes en diferentes áreas, mientras que técnicas de agrupamiento pueden segmentar a los estudiantes según características similares. (EducaOpen, s.f.)

1.2 Descripción de caso de estudio.

Para nuestro análisis, utilizaremos el conjunto de datos 'Student Performance-BD'. Esta base de datos, compuesta por 24 variables, nos permitirá explorar en profundidad los factores que influyen en el rendimiento académico de estudiantes de diferentes regiones de Bangladesh entre las que destacan el promedio de calificaciones en cada asignatura, el nivel socioeconómico de los estudiantes, el tamaño de la familia, la participación en actividades extracurriculares y el acceso a recursos tecnológicos. Utilizaremos este conjunto de datos para identificar los factores que más influyen en el rendimiento académico de los estudiantes bangladesíes, con el objetivo de diseñar estrategias para mejorar la calidad de la educación

1.3 Pregunta problema:

¿Cómo desarrollar una estrategia computacional para describir el rendimiento académico de los estudiantes bangladesíes, y poder a partir de estos, realizar mejoras, haciendo uso de algoritmos de Machine Learning?

1.4 Hipótesis:

El análisis computacional de los datos de puntajes de los estudiantes bangladesíes en 5 áreas fundamentales, y la conexión al conocimiento, son relevantes, para hacer un seguimiento de capacidades, y así visualizar falencias y virtudes dentro de los campos evaluados, poniendo en contexto mejoras que se puedan llevar a cabo, para aumentar el grado de conocimiento de estos estudiantes.

2. Objetivos

2.1 Objetivo general.

Analizar el rendimiento académico de los estudiantes bangladesíes, en las áreas de Matemáticas, inglés, Ciencia, Ciencias Sociales y Arte y cultura, a través del tratamiento y modelado de datos, empleando algoritmos de estadística descriptiva y modelos de regresión predictiva, con el fin de identificar fortalezas y debilidades en dichas áreas y proponer estrategias de mejora.

2.2 Objetivos específicos.

- Analizar el rendimiento académico de los estudiantes bangladesíes en las áreas de Matemáticas, inglés, ciencia, ciencias sociales, arte y cultura a través de los datos proporcionados en la base estudiantil.

- Implementar algoritmos de tratamiento de datos para la eliminación de datos no relevantes y modificar los datos categóricos y alfabéticos a datos numéricos, para el análisis de los datos.

- Aplicar modelos de regresión y predicción, para calcular los promedios estimados y analizar los márgenes de error.

- Analizar el rendimiento académico de los estudiantes bangladesíes, en las asignaturas de Matemáticas, inglés, ciencia, ciencias sociales, arte y cultura.

- Proporcionar recomendaciones basadas en las estadísticas obtenidas, orientadas a mejorar el rendimiento académico de los estudiantes bangladesíes.

-

3. Desarrollo e implementación del aprendizaje

En este trabajo se implementa el modelo de modificación de la base de datos de puntajes de estudiantes bangladesíes, para proceder a ejecutar algoritmos que me generen gráficas estadísticas, analizando los resultados e implementando otros algoritmos de modelos de regresión tales como el Knn y el Ann, que se mostraran a continuación.

3.1 Preparación y análisis de los datos

Se escoge el dataset de estudio Rendimiento estudiantil-BD, en esta base de datos se concentran los datos de los estudiantes de educación media de la ciudad de bangladesh, la cual contiene el rendimiento académico de los estudiantes los cuales se recopilaron de varias instituciones, tanto públicas como privadas, donde dan datos relevantes para el estudio que escogimos.

3.1.1 carga de datos

```
[12] #Para cargar los datos
     import pandas as pd
     from google.colab import files
     uploaded = files.upload()
     for filename in uploaded.keys():
         Conjunto_Datos = pd.read_csv(filename,sep=',')
         #Conjunto_Datos = pd.read_excel(filename)
     Conjunto_Datos.head()
```

Choose Files bd_students_per_v2.csv

- bd_students_per_v2.csv(text/csv) - 980259 bytes, last modified: 8/12/2024 - 100% done

Saving bd_students_per_v2.csv to bd_students_per_v2.csv

	id	full_name	age	gender	location	family_size	mother_education	father_education	mother_job	father_job	...	tutoring	school_type	attendance	extra_curricular_activities	english	math	science	social_science	art_culture	stu_group
0	2	Avi Biswas	16	Male	Urban	6	SSC	HSC	No	No	...	Yes	Private	95	Yes	95	98	92	94	98	Science
1	3	Taslima Sultana	18	Female	Rural	6	SSC	HSC	No	Yes	...	No	Semi_Govt	92	No	65	71	40	78	80	Commerce
2	4	Md Adilur Rahman	15	Male	Rural	4	SSC	SSC	Yes	Yes	...	Yes	Govt	81	Yes	64	78	58	86	74	Commerce
3	5	Saleh Ahmed	16	Male	Rural	6	SSC	SSC	Yes	Yes	...	Yes	Private	90	Yes	84	90	85	86	88	Science
4	6	Din Islam	17	Male	Urban	5	Honors	Masters	No	Yes	...	Yes	Semi_Govt	75	Yes	54	70	45	79	76	Commerce

5 rows × 24 columns

Para iniciar con el análisis de datos, cargamos el dataset bd_students.csv, se hace la carga de datos utilizando un algoritmo para este propósito.

3.1.2. Identificación de las columnas

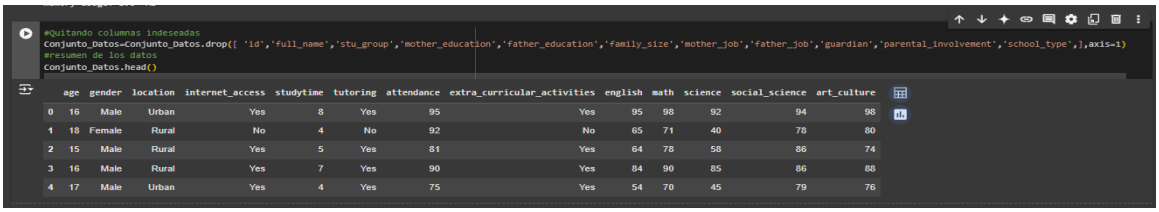
```
#Información de la estructura de datos
Conjunto_Datos.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 8612 entries, 0 to 8611
Data columns (total 24 columns):
 #   Column                     Non-Null Count  Dtype 
---  ------                     --------------  ----- 
 0   id                         8612 non-null   int64 
 1   full_name                  8612 non-null   object
 2   age                        8612 non-null   int64 
 3   gender                     8612 non-null   object
 4   location                   8611 non-null   object
 5   family_size                8612 non-null   int64 
 6   mother_education           8602 non-null   object
 7   father_education           8608 non-null   object
 8   mother_job                 8612 non-null   object
 9   father_job                 8612 non-null   object
 10  guardian                   8612 non-null   object
 11  parental_involvement       8612 non-null   object
 12  internet_access            8612 non-null   object
 13  studytime                  8612 non-null   int64 
 14  tutoring                   8612 non-null   object
 15  school_type                8612 non-null   object
 16  attendance                 8612 non-null   int64 
 17  extra_curricular_activities  8612 non-null   object
 18  english                    8612 non-null   int64 
 19  math                       8612 non-null   int64 
 20  science                    8612 non-null   int64 
 21  social_science             8612 non-null   int64 
 22  art_culture                8612 non-null   int64 
 23  stu_group                  8612 non-null   object
dtypes: int64(10), object(14)
memory usage: 1.6+ MB
```

Analizamos las columnas desplegadas para obtener un panorama mas claro de las columnas que vamos a utilizar para este analisis, clasificando la importancia de los datos contenidos para el estudio.

3.1.3 Modelado de los datos

3.1.3.1 Borrado de datos que no se tendrán en cuentan

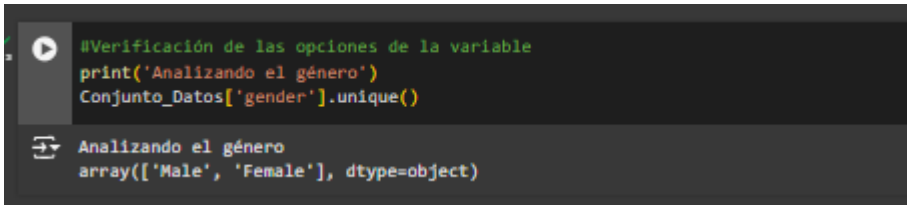


```
#Quitando columnas indeseadas
Conjunto_Datos=Conjunto_Datos.drop([ 'id','full_name','stu_group','mother_education','father_education','family_size','mother_job','father_job','guardian','parental_involvement','school_type',],axis=1)
#resumen de los datos
conjunto_datos.head()
```

	age	gender	location	internet_access	studytime	tutoring	attendance	extra_curricular_activities	english	math	science	social_science	art_culture
0	16	Male	Urban	Yes	8	Yes	95	Yes	95	98	92	94	98
1	18	Female	Rural	No	4	No	92	No	65	71	40	78	80
2	15	Male	Rural	Yes	5	Yes	81	Yes	64	78	58	86	74
3	16	Male	Rural	Yes	7	Yes	90	Yes	84	90	85	86	88
4	17	Male	Urban	Yes	4	Yes	75	Yes	54	70	45	79	76

Eliminamos las columnas que no vamos a utilizar, quitándolas de la carga de datos que tenemos, para este fin eliminamos las siguientes columnas: 'id','full_name','stu_group','mother_education','father_education','family_size','mother_job','father_job','guardian','parental_involvement','school_type' , stu group dejando los datos de mayor relevancia como se muestra la gráfica anterior

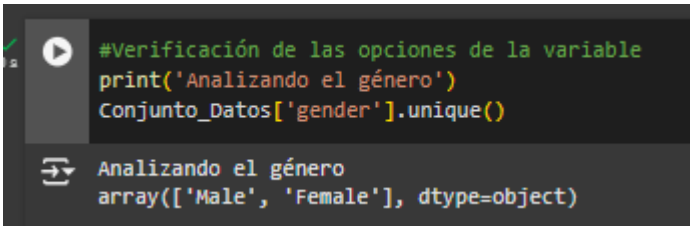
3.1.3.2 cambio de datos alfabéticos a numéricos



```
#Verificación de las opciones de la variable
print('Analizando el género')
Conjunto_Datos['gender'].unique()
```

```
Analizando el género
array(['Male', 'Female'], dtype=object)
```

Se verifica la primera variable, la cual es gender, verificamos los parámetros en la columna, los datos de la columna son alfabéticos, para mejorar la recepción de los datos y hacer los cálculos correspondientes, se modifican los datos para dejarlos numéricos



```
#Verificación de las opciones de la variable
print('Analizando el género')
Conjunto_Datos['gender'].unique()
```

```
Analizando el género
array(['Male', 'Female'], dtype=object)
```

Se realiza el cambio de los datos del campo género, en parámetros numéricos de la siguiente forma, (hombre 1),(mujer 2.).

```
Reemplazo_1={'Male':1,'Female':2}
Conjunto_Datos['gender']=Conjunto_Datos['gender'].map(Reemplazo_1)

Conjunto_Datos.head()
```

	age	gender	location	internet_access	studytime	tutoring	attendance	extra_curricular_activities	english	math	science	social_science	art_culture
0	16	1	Urban	Yes	8	Yes	95	Yes	95	98	92	94	98
1	18	2	Rural	No	4	No	92	No	65	71	40	78	80
2	15	1	Rural	Yes	5	Yes	81	Yes	64	78	58	86	74
3	16	1	Rural	Yes	7	Yes	90	Yes	84	90	85	86	88
4	17	1	Urban	Yes	4	Yes	75	Yes	54	70	45	79	76

Se procede con el cambio y se deja reflejado en la columna

```
Reemplazo_1={'Urban':1,'urban':2,'City':3,'city':4,'Rural':5}
Conjunto_Datos['location']=Conjunto_Datos['location'].map(Reemplazo_1)

Conjunto_Datos.head()
```

	age	gender	location	internet_access	studytime	tutoring	attendance	extra_curricular_activities	english	math	science	social_science	art_culture
0	16	1	1	Yes	8	Yes	95	Yes	95	98	92	94	98
1	18	2	5	No	4	No	92	No	65	71	40	78	80
2	15	1	5	Yes	5	Yes	81	Yes	64	78	58	86	74
3	16	1	5	Yes	7	Yes	90	Yes	84	90	85	86	88
4	17	1	1	Yes	4	Yes	75	Yes	54	70	45	79	76

Como en el anterior caso, la columna location también tiene datos alfabéticos y se reemplazan los datos de la siguiente manera 'Urban':1,'urban':2,'City':3,'city':4,'Rural':5

Se numeran de forma continua ya que los datos no son homogéneos se ubican los datos, de forma consecutiva, para dejar los datos ordenados

```
[36]
print('Analizando internet_access')
Conjunto_Datos['internet_access'].unique()

Analizando internet_access
array(['Yes', 'No'], dtype=object)
```

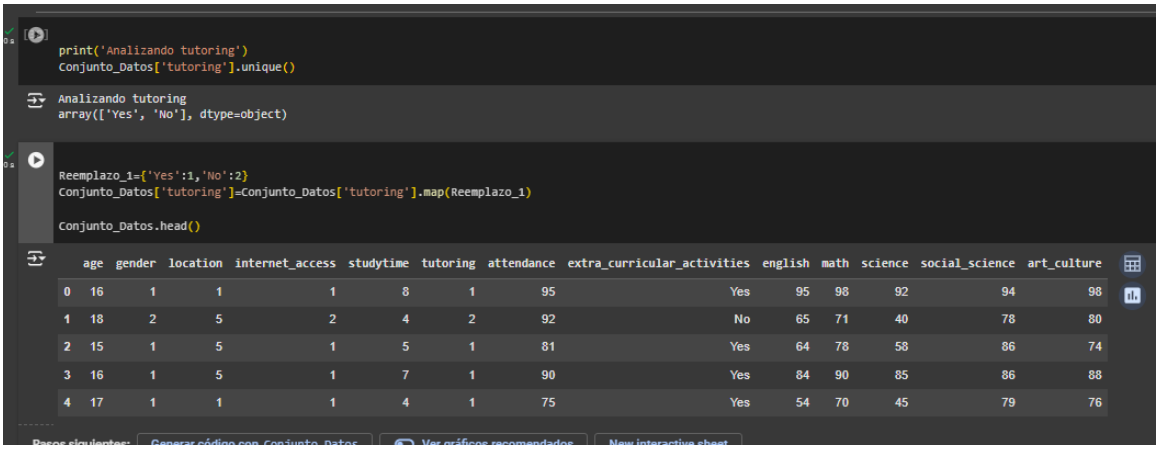
En la variable internet access se hace el mismo proceso que en los parámetros anteriores para ir modelando los datos numéricos, esto con el fin de que al momento de usar los datos no hayan inconvenientes con el manejo de los datos.

En los siguientes parámetros se hace el mismo proceso que el anterior para dejar la tabla con datos numéricos para empezar a generar las gráficas para el análisis.

```
Reemplazo_1={'Yes':1,'No':2}
Conjunto_Datos['internet_access']=Conjunto_Datos['internet_access'].map(Reemplazo_1)

Conjunto_Datos.head()
```

	age	gender	location	internet_access	studytime	tutoring	attendance	extra_curricular_activities	english	math	science	social_science	art_culture
0	16	1	1	1	8	Yes	95	Yes	95	98	92	94	98
1	18	2	5	2	4	No	92	No	65	71	40	78	80
2	15	1	5	1	5	Yes	81	Yes	64	78	58	86	74
3	16	1	5	1	7	Yes	90	Yes	84	90	85	86	88
4	17	1	1	1	4	Yes	75	Yes	54	70	45	79	76

Se cambian los datos de la siguiente manera 'Yes':1,'No':2.



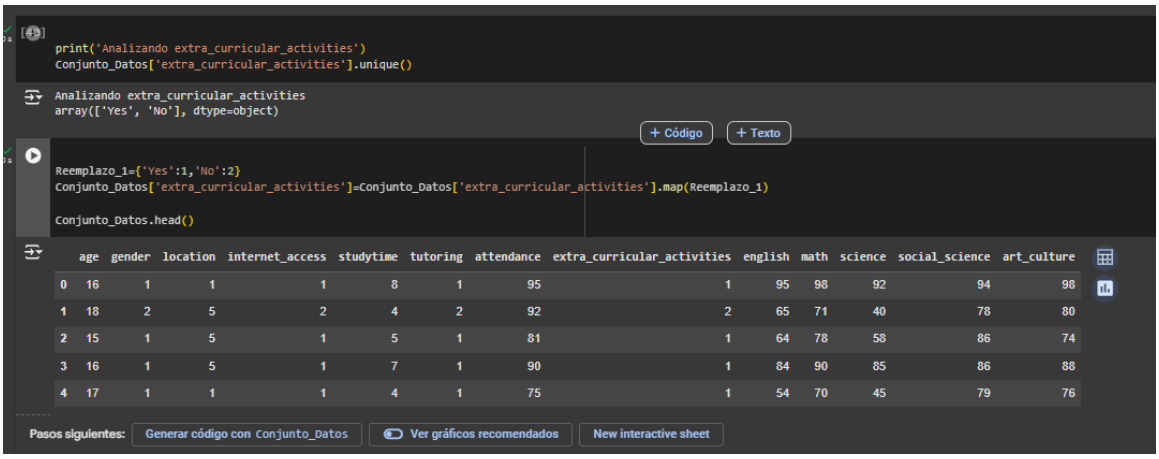
```
print('Analizando tutoring')
Conjunto_Datos['tutoring'].unique()
```

```
Analizando tutoring
array(['Yes', 'No'], dtype=object)
```

```
Reemplazo_1={'Yes':1,'No':2}
Conjunto_Datos['tutoring']=Conjunto_Datos['tutoring'].map(Reemplazo_1)

Conjunto_Datos.head()
```

	age	gender	location	internet_access	studytime	tutoring	attendance	extra_curricular_activities	english	math	science	social_science	art_culture
0	16	1	1	1	8	1	95	Yes	95	98	92	94	98
1	18	2	5	2	4	2	92	No	65	71	40	78	80
2	15	1	5	1	5	1	81	Yes	64	78	58	86	74
3	16	1	5	1	7	1	90	Yes	84	90	85	86	88
4	17	1	1	1	4	1	75	Yes	54	70	45	79	76

se cambian los datos de la siguiente manera 'Yes':1,'No':2



```
print('Analizando extra_curricular_activities')
Conjunto_Datos['extra_curricular_activities'].unique()
```

```
Analizando extra_curricular_activities
array(['Yes', 'No'], dtype=object)
```

+ Código + Texto

```
Reemplazo_1={'Yes':1,'No':2}
Conjunto_Datos['extra_curricular_activities']=Conjunto_Datos['extra_curricular_activities'].map(Reemplazo_1)

Conjunto_Datos.head()
```

	age	gender	location	internet_access	studytime	tutoring	attendance	extra_curricular_activities	english	math	science	social_science	art_culture
0	16	1	1	1	8	1	95	1	95	98	92	94	98
1	18	2	5	2	4	2	92	2	65	71	40	78	80
2	15	1	5	1	5	1	81	1	64	78	58	86	74
3	16	1	5	1	7	1	90	1	84	90	85	86	88
4	17	1	1	1	4	1	75	1	54	70	45	79	76

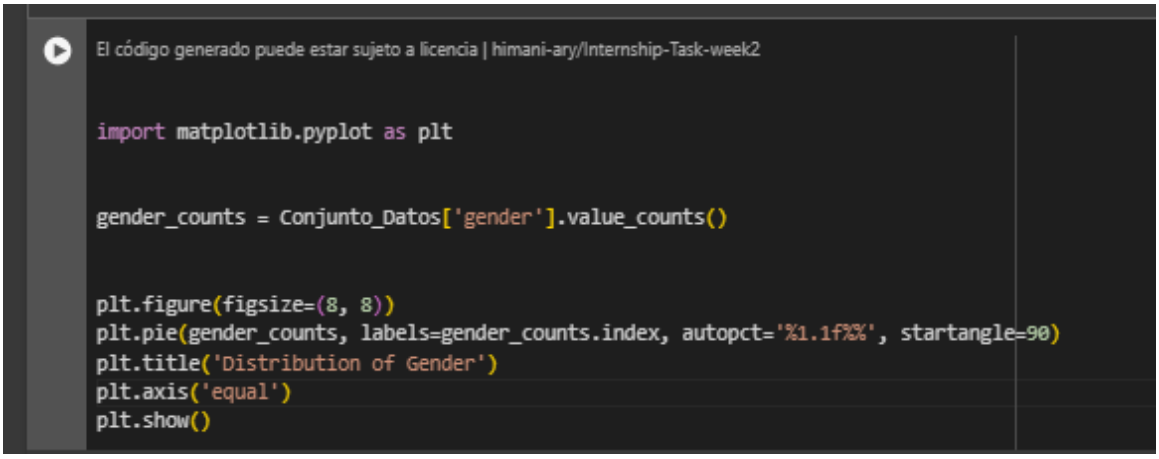
Pasos siguientes: Generar código con Conjunto_Datos | Ver gráficos recomendados | New interactive sheet

Se cambian los datos de la siguiente manera 'Yes':1,'No':2.

Se termina el proceso haciendo una validación y borrado de espacios en blanco para que no afecten los resultados de los modelos a utilizar,

3.2 Modelo de toma de decisiones

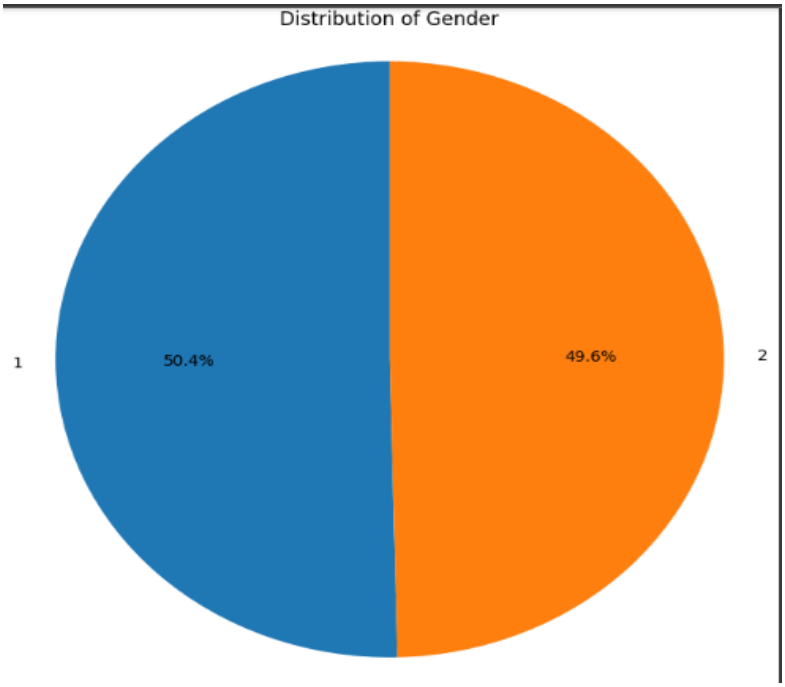
Puede ser clustering o lo que hace falta por ver. Todo lo que hicimos en el resto de las clases, se debe explicar desde el punto de vista de ustedes, pegando gráficos y pantallazos de resultados.



El código generado puede estar sujeto a licencia | himani-ary/Internship-Task-week2

```
import matplotlib.pyplot as plt


gender_counts = Conjunto_Datos['gender'].value_counts()


plt.figure(figsize=(8, 8))
plt.pie(gender_counts, labels=gender_counts.index, autopct='%1.1f%%', startangle=90)
plt.title('Distribution of Gender')
plt.axis('equal')
plt.show()
```

Distribución de la población estudiantil por género

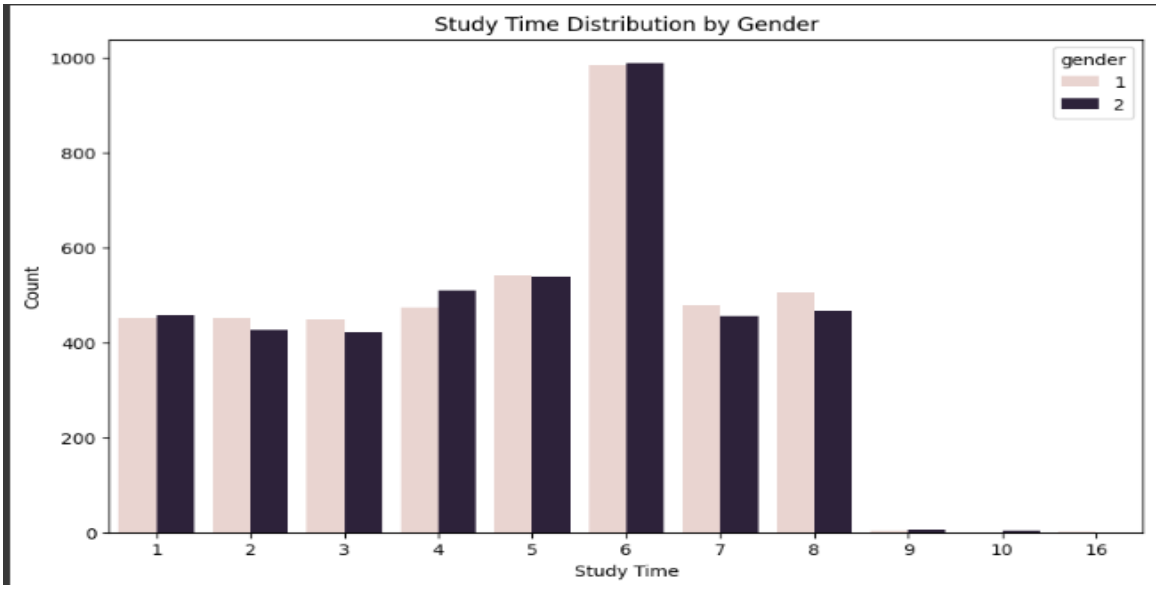


En la gráfica podemos apreciar el acceso a la educación por géneros donde nos muestra un 50.4 por ciento para hombres y un 49.6 por ciento para las mujeres, podemos concluir que el acceso de los estudiantes es equitativo

```
import matplotlib.pyplot as plt
import seaborn as sns

plt.figure(figsize=(10, 6))
sns.countplot(x='studytime', hue='gender', data=Conjunto_Datos)
plt.title('Study Time Distribution by Gender')
plt.xlabel('Study Time')
plt.ylabel('Count')
plt.show()
```

Tiempo dedicado al estudio por géneros.

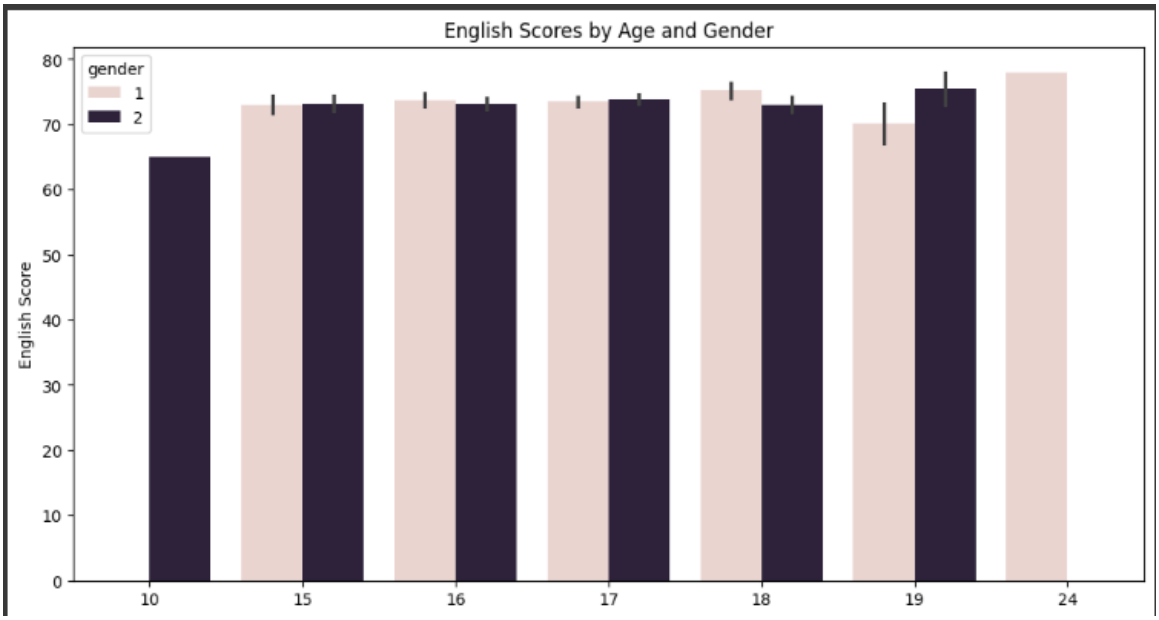
Gráfica de distribución de estudio



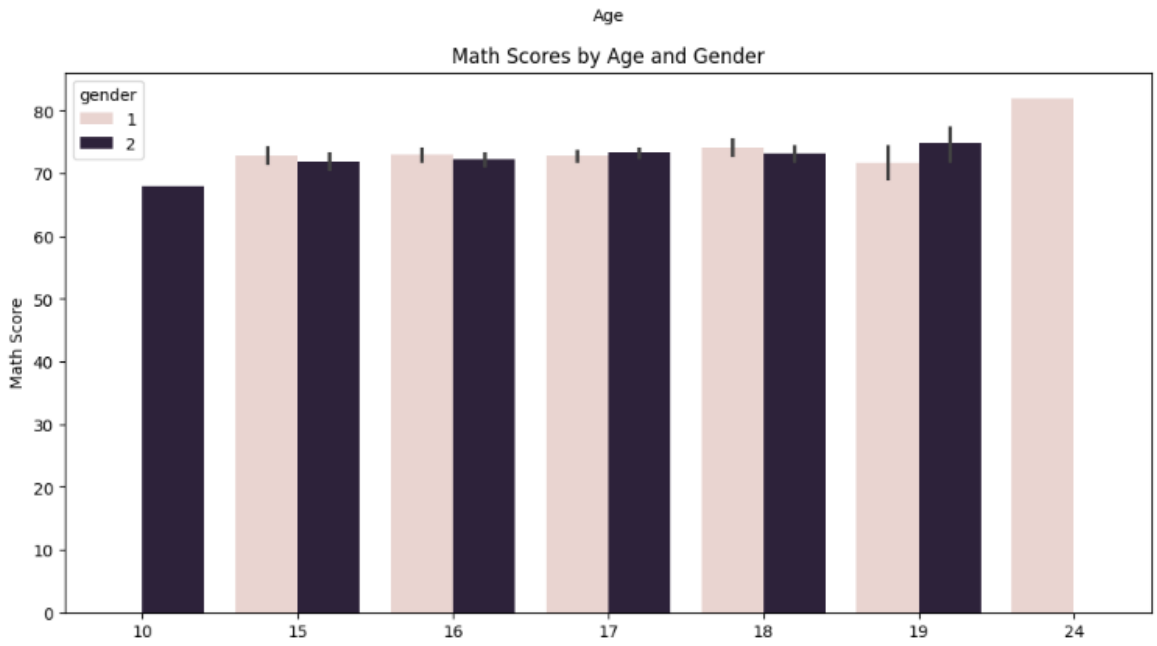
El tiempo que dedican los estudiantes al estudio por fuera de las aulas de clase, los estudiantes tanto masculinos como femeninos dedican en su mayoría 6 horas de estudio, siguiendo de una gran cantidad que dedica 5 horas de estudio al día.

```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns


plt.figure(figsize=(12, 6))
sns.barplot(x='age', y='english', hue='gender', data=Conjunto_Datos)
plt.title('English Scores by Age and Gender')
plt.xlabel('Age')
plt.ylabel('English Score')
plt.show()

plt.figure(figsize=(12, 6))
sns.barplot(x='age', y='math', hue='gender', data=Conjunto_Datos)
plt.title('Math Scores by Age and Gender')
plt.xlabel('Age')
plt.ylabel('Math Score')
plt.show()

plt.figure(figsize=(12, 6))
sns.barplot(x='age', y='science', hue='gender', data=Conjunto_Datos)
plt.title('Science Scores by Age and Gender')
plt.xlabel('Age')
plt.ylabel('Science Score')
plt.show()
```

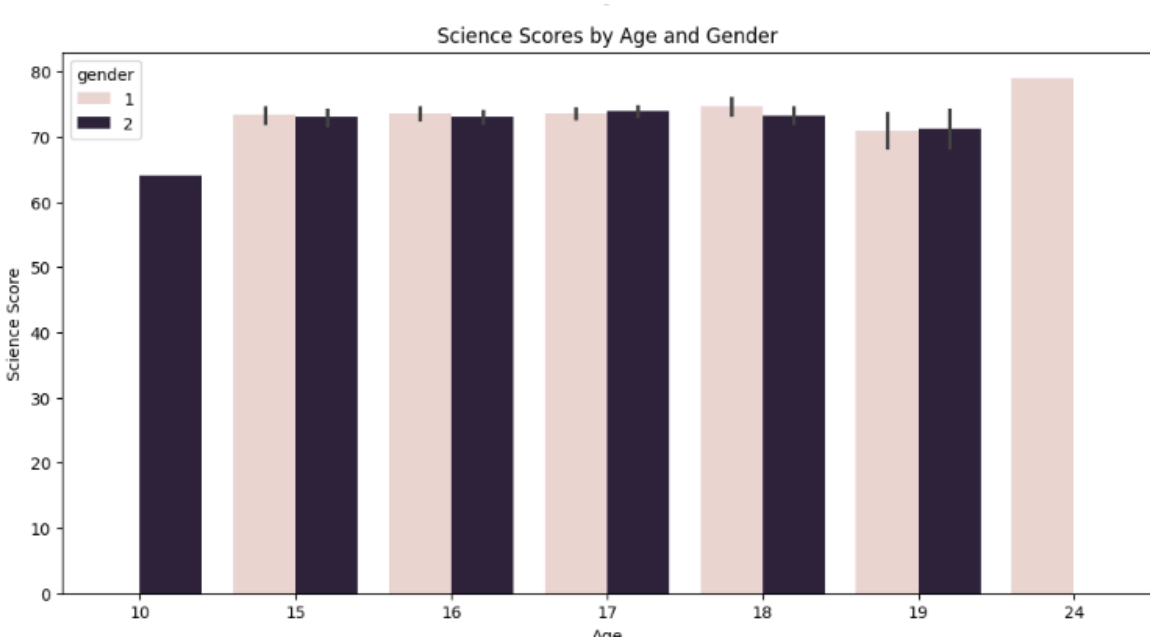
Notas en inglés



Notas en matemáticas



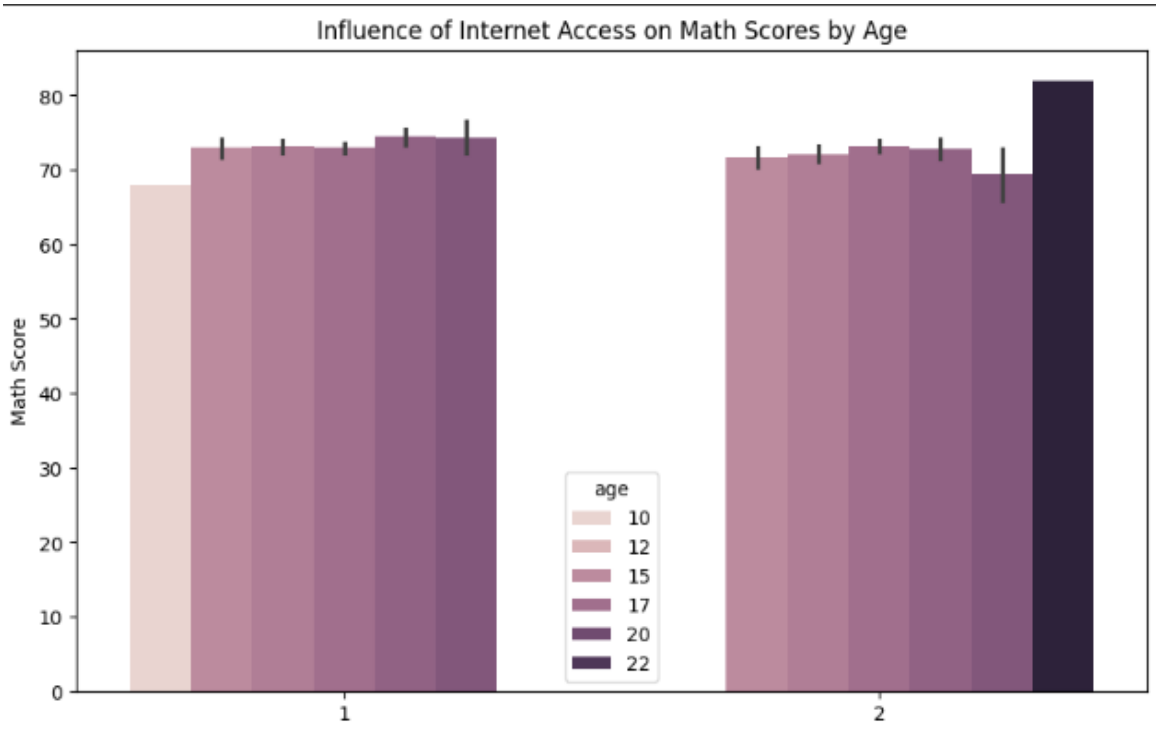
Notas en Ciencia

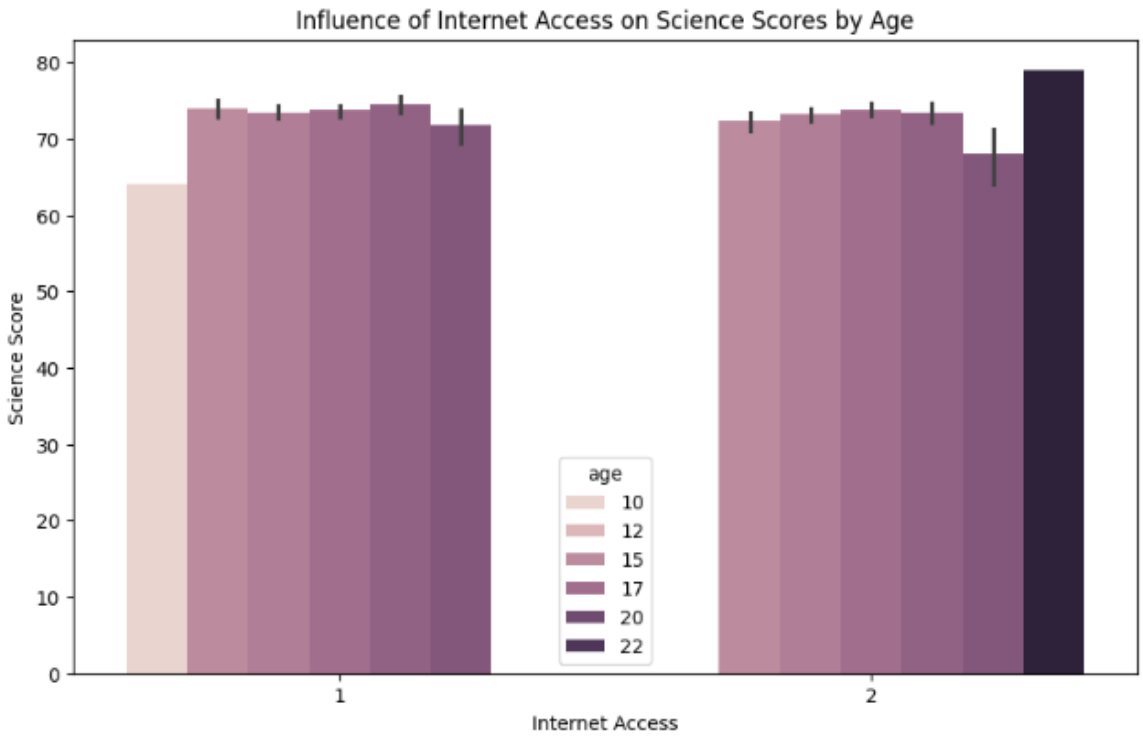


La tendencia en la educación es que las mujeres inician los estudios a más temprana edad, y los hombres según la tendencia en promedio se gradúan con una edad más adulta y el rendimiento es muy parejo hasta las materias de inglés y matemáticas donde las mujeres sobresalen académicamente a la edad de 19 años, por el caso masculino es a los 18 años, las variables en las otras edades se mantienen equitativas en los resultados.

Acceso a internet

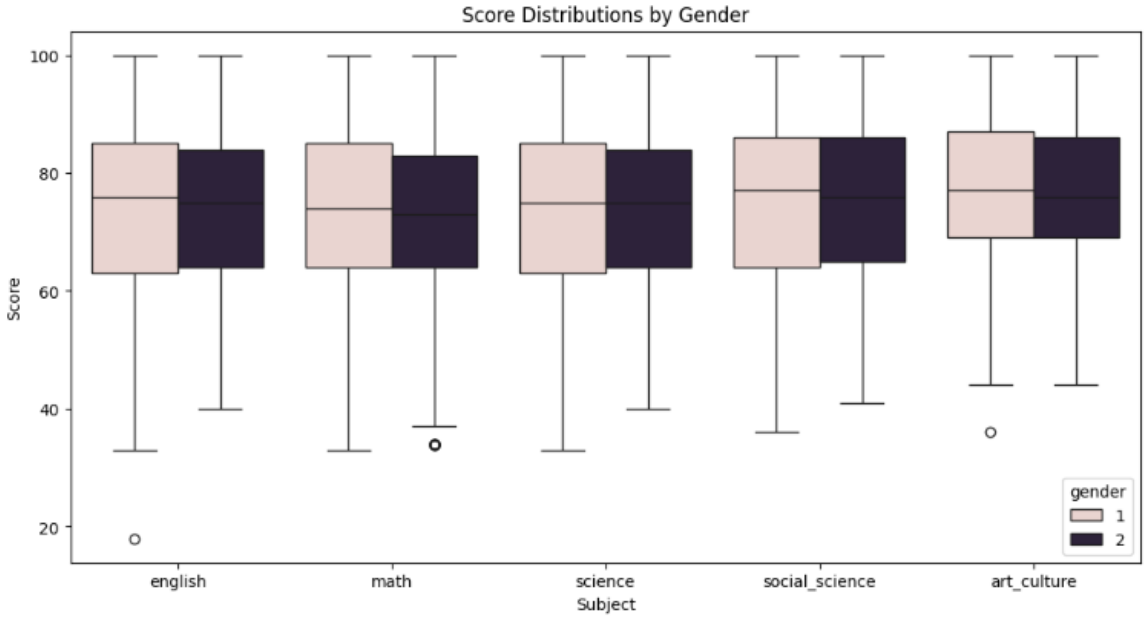
```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns

plt.figure(figsize=(10, 6))
sns.barplot(x='internet_access', y='math', hue='age', data=Conjunto_Datos)
plt.title('Influence of Internet Access on Math Scores by Age')
plt.xlabel('Internet Access')
plt.ylabel('Math Score')
plt.show()


plt.figure(figsize=(10, 6))
sns.barplot(x='internet_access', y='science', hue='age', data=Conjunto_Datos)
plt.title('Influence of Internet Access on Science Scores by Age')
plt.xlabel('Internet Access')
plt.ylabel('Science Score')
plt.show()
```





En la gráfica podemos apreciar como el uso de la tecnología de la información impacta directamente en la calidad de estudio y el rendimiento académico de los estudiantes donde en la gráfica en la parte 1 son los estudiantes que tienen acceso a internet y en el segundo lado son los que no tienen un acceso. la diferencia incrementa de acuerdo al dominio de la tecnología y la edad del estudiante juega un papel importante en el rendimiento académico general, en particular analizando los datos de matemáticas, la nota promedio más alta sin acceso a internet es la nota promedio de las personas que tienen acceso, en lo que podemos resumir que la herramienta ayuda a mejorar significativamente el rendimiento de los estudiantes.

Relación de los promedios por materia



Analizamos el promedio general entre estudiantes masculinos y femeninos y observamos que el estudiante masculino tiene un promedio en calificaciones más alto que las mujeres en el área de mayor diferencia son las matemáticas y la más equitativa son ciencias sociales arte y cultura, los promedios de las materias se sostienen entre 70 y 80 y la materia donde el promedio es igual es en ciencias.

3.3 Análisis de desempeño

En esta parte mostramos cómo se comporta el algoritmo. Todo lo que hicimos en el resto de las clases, se debe explicar desde el punto de vista de ustedes, pegando gráficos y pantallazos de resultados.

3.3.1 Modelo KNN

```python
import pandas as pd
import numpy as np
from sklearn.neighbors import KNeighborsRegressor as KNNR
from sklearn.model_selection import train_test_split

np.random.seed(42)
Mis_datos = pd.DataFrame({
    'english': np.random.randint(50, 100, 100),
    'math': np.random.randint(50, 100, 100),
    'science': np.random.randint(50, 100, 100),
    'social_science': np.random.randint(50, 100, 100),
    'art_culture': np.random.randint(50, 100, 100),
    'porcentaje': np.random.randint(50, 100, 100)
})

X = Mis_datos[['english', 'math', 'science', 'social_science', 'art_culture']]
y = Mis_datos['porcentaje']


X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

Modelo_1 = KNNR()
Modelo_1.fit(X_train, y_train)

Nueva_entrada = np.zeros((1, 5))
Nueva_entrada[0, 0] = float(input('Ingrese la calificación de inglés: '))
Nueva_entrada[0, 1] = float(input('Ingrese la calificación de matemáticas: '))
Nueva_entrada[0, 2] = float(input('Ingrese la calificación de ciencias: '))
Nueva_entrada[0, 3] = float(input('Ingrese la calificación de ciencias sociales: '))
Nueva_entrada[0, 4] = float(input('Ingrese la calificación de arte y cultura: '))

# Predicción
try:
    Proyeccion_1 = np.round(Modelo_1.predict(Nueva_entrada))
    print('Según los datos ingresados, la proyección del porcentaje usando KNN será: ', Proyeccion_1[0])
except ValueError as e:
    print(f"Error: {e}. Asegúrese de que la forma de la entrada coincida con la forma de los datos de entrenamiento.")
```

Resultado modelo KNN

```
Ingrese la calificación de inglés: 80
Ingrese la calificación de matemáticas: 90
Ingrese la calificación de ciencias: 80
Ingrese la calificación de ciencias sociales: 80
Ingrese la calificación de arte y cultura: 80
Según los datos ingresados, la proyección del porcentaje usando KNN será:  71.0
/usr/local/lib/python3.10/dist-packages/sklearn/base.py:493: UserWarning: X does not have valid feature names, but KNeighborsRegressor was fitted with feature names
  warnings.warn(
```

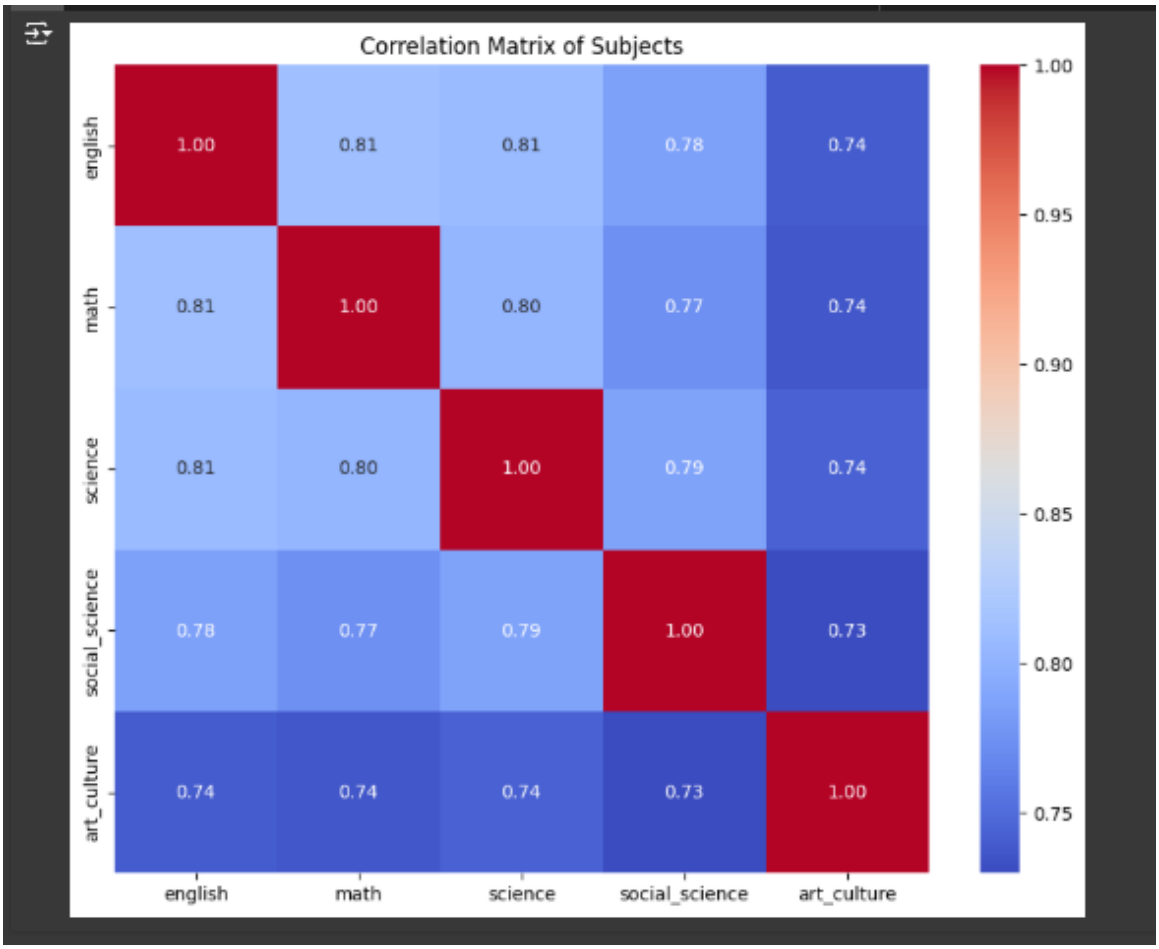
3.3.2 Modelo ANN

```
import numpy as np
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler
from tensorflow.keras.models import Sequential
from tensorflow.keras.layers import Dense

np.random.seed(42)
Mis_datos = pd.DataFrame({
    'english': np.random.randint(50, 100, 100),
    'math': np.random.randint(50, 100, 100),
    'science': np.random.randint(50, 100, 100),
    'social_science': np.random.randint(50, 100, 100),
    'art_culture': np.random.randint(50, 100, 100),
    'porcentaje': np.random.randint(50, 100, 100)
})

X = Mis_datos[['english', 'math', 'science', 'social_science', 'art_culture']]
y = Mis_datos['porcentaje']

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

scaler = StandardScaler()
X_train_scaled = scaler.fit_transform(X_train)
X_test_scaled = scaler.transform(X_test)

model = Sequential([
    Dense(64, activation='relu', input_shape=(X_train.shape[1],)),
    Dense(32, activation='relu'),
    Dense(1)
])

model.compile(optimizer='adam', loss='mse', metrics=['mae'])

history = model.fit(X_train_scaled, y_train, epochs=50, batch_size=8, validation_data=(X_test_scaled, y_test), verbose=1)

Nueva_entrada = np.zeros((1, 5))
Nueva_entrada[0, 0] = float(input('Ingrese la calificación de inglés: '))
Nueva_entrada[0, 1] = float(input('Ingrese la calificación de matemáticas: '))
Nueva_entrada[0, 2] = float(input('Ingrese la calificación de ciencias: '))
Nueva_entrada[0, 3] = float(input('Ingrese la calificación de ciencias sociales: '))
Nueva_entrada[0, 4] = float(input('Ingrese la calificación de arte y cultura: '))
```

Resultado modelo ANN

```
Ingrese la calificación de inglés: 79
Ingrese la calificación de matemáticas: 81
Ingrese la calificación de ciencias: 84
Ingrese la calificación de ciencias sociales: 88
Ingrese la calificación de arte y cultura: 90
1/1 ━━━━━━━━━━━━━━━━━━━━ 0s 63ms/step
Según los datos ingresados, la proyección del porcentaje usando ANN será:  64.19517
```

3.3.3 Matriz de correlación



Las materias más dominantes entre los estudiantes son matemáticas e inglés y las que menos dominan son ciencias sociales, arte y cultura.

3.4 Validación de los modelos

Base de datos estudiantes "Este dataset tiene como objetivo analizar los puntajes de los estudiantes en las asignaturas Matemáticas, Inglés, Ciencia, Ciencias Sociales, Arte y cultura. Las variables involucradas son:

Relación de las variables

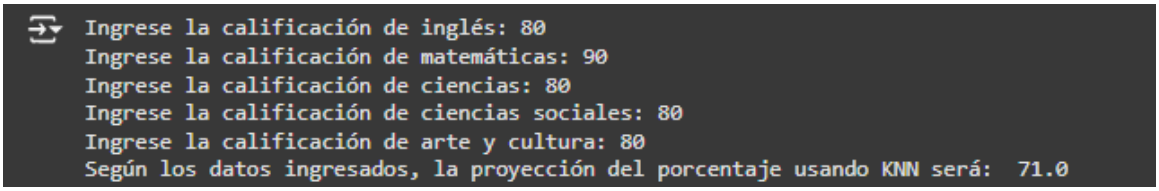
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 8612 entries, 0 to 8611
Data columns (total 13 columns):
 #   Column                      Non-Null Count  Dtype 
---  ------                      --------------  ----- 
 0   age                         8612 non-null   int64 
 1   gender                      8612 non-null   object
 2   location                    8611 non-null   object
 3   internet_access             8612 non-null   object
 4   studytime                   8612 non-null   int64 
 5   tutoring                    8612 non-null   object
 6   attendance                  8612 non-null   int64 
 7   extra_curricular_activities 8612 non-null   object
 8   english                     8612 non-null   int64 
 9   math                        8612 non-null   int64 
 10  science                     8612 non-null   int64 
 11  social_science              8612 non-null   int64 
 12  art_culture                 8612 non-null   int64 
dtypes: int64(8), object(5)
memory usage: 874.8+ KB
```

Descripción de las variables

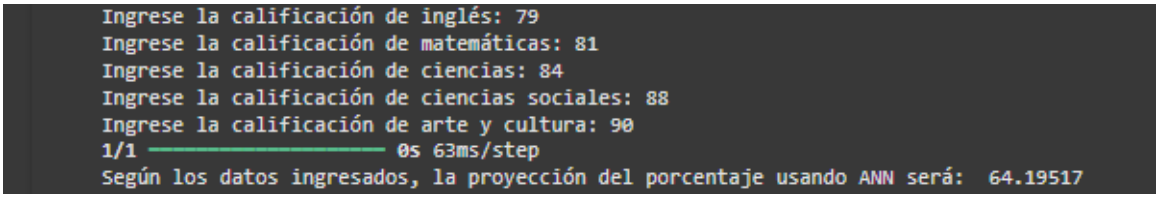
```
Conjunto_Datos.describe()
```

	age	studytime	attendance	english	math	science	social_science	art_culture
count	8612.000000	8612.000000	8612.000000	8612.000000	8612.000000	8612.000000	8612.000000	8612.000000
mean	16.614375	4.746168	74.027055	73.610427	72.938110	73.524036	74.465513	76.236066
std	0.969057	2.185645	13.290046	15.343903	15.816147	15.115104	14.945846	13.619125
min	10.000000	1.000000	30.000000	18.000000	33.000000	33.000000	36.000000	36.000000
25%	16.000000	3.000000	63.000000	64.000000	64.000000	63.000000	65.000000	69.000000
50%	17.000000	5.000000	73.000000	75.000000	73.000000	75.000000	77.000000	76.000000
75%	17.000000	6.000000	86.000000	85.000000	84.000000	85.000000	86.000000	86.000000
max	24.000000	16.000000	100.000000	100.000000	100.000000	100.000000	100.000000	100.000000

Resultado del modelo KNN



Resultado del Modelo ANN



Usando los modelos KNN y ANN podemos visualizar los márgenes de error de unos determinados puntajes a evaluar, y cual sería su media ponderada o promedio.

4. Conclusiones y trabajos futuros

En la actualidad, los datos se han convertido en uno de los activos más valiosos, pues abren un abanico infinito de posibilidades. Sin embargo, en los países en desarrollo, la recolección y gestión de información a menudo se subestiman. La falta de sistemas confiables y eficientes para almacenar y analizar datos nos impide aprovechar su potencial para mejorar procesos clave, como la educación. Es fundamental reconocer que la información detallada sobre los estudiantes y los sistemas educativos puede revelar patrones, identificar deficiencias y orientar la toma de decisiones para optimizar los resultados académicos.

Se puede observar que en los análisis se ve la diferencia de puntajes entre los estudiantes que tienen acceso a internet y los que no lo tienen, resaltando la importancia de este recurso a nivel educativo.

Se puede decir que la educación es muy equitativa, pues se puede ver casi un 50/50 de estudiantes entre hombres y mujeres, y aunque se tienen casi las mismas horas de estudio entre hombres y mujeres, se puede apreciar que los hombres presentan un mejor rendimiento académico que las mujeres.

Se observa que, en las materias, a medida que aumenta la edad de los estudiantes, se empieza a notar que el rendimiento de las mujeres aumenta frente al de los hombres.

Analizando los modelos KNN y ANN, podemos realizar predicciones de puntajes que se ingresen para saber la calificación que se espera por parte de un último estudiante, que podría llevarnos a saber cuál será el próximo puntaje que se vería en la base de datos, y con base en esa predicción, se sabrá que decisiones tomar a futuro sobre estas materias y su enseñanza.

Referencias bibliográficas

https://www.educaopen.com/digital-lab/blog/educacion-digital/learning-analytics

https://www.ellucian.com/es/ideas/como-usar-el-analisis-de-datos-para-incrementar-el-exito-estudiantil

https://www.iberdrola.com/innovacion/machine-learning-aprendizaje-automatico

https://www.ibm.com/mx-es/topics/knn

https://learn.microsoft.com/es-es/azure/cosmos-db/gen-ai/knn-vs-ann

https://www.kaggle.com/datasets/satayjit/student-performance-bd/data