Proyecto de curso

Dennis Adriana González Cifuentes
Ciencias Naturales e Ingeniería
Universidad Jorge Tadeo Lozano
Bogotá - Colombia
dennisa.gonzalezc@utadeo.edu.co

María Fernanda López Martínez
Ciencias Naturales e Ingeniería
Universidad Jorge Tadeo Lozano
Bogotá - Colombia
mariafe.lopezm@utadeo.edu.co

Resumen

En el presente documento se encuentra el proyecto de Inteligencia Artificial, en el cual se presentará la extracción, pre-procesamiento, visualización y análisis de los datos escogidos previamente *(Notas obtenidas por estudiantes en varias asignaturas)*, el cual se realizará por medio de la metodología y las herramientas presentadas en el curso.

Palabras cláve

Notas, asignaturas, calificaciones, estudiantes, Estados Unidos, grupos étnicos, matemáticas, lectura, escritura.

1. Marco teórico

La recolección de notas obtenidas de estudiantes de secundaria en un colegio ubicado en Estados Unidos, nos brindan todo tipo de información la cual será realizada e intrepretada en el presente documento, con el fin de encontrar y visualizar las principales características estadísticas de estos utilizando las herramientas vistas en clases

2. Visualización

Para empezar se deben importar las librerías adecuadas para la realización del trabajo, el cual se escogío un conjunto de datos de las *Notas obtenidas por estudiantes en varias asignaturas de secuandaria*

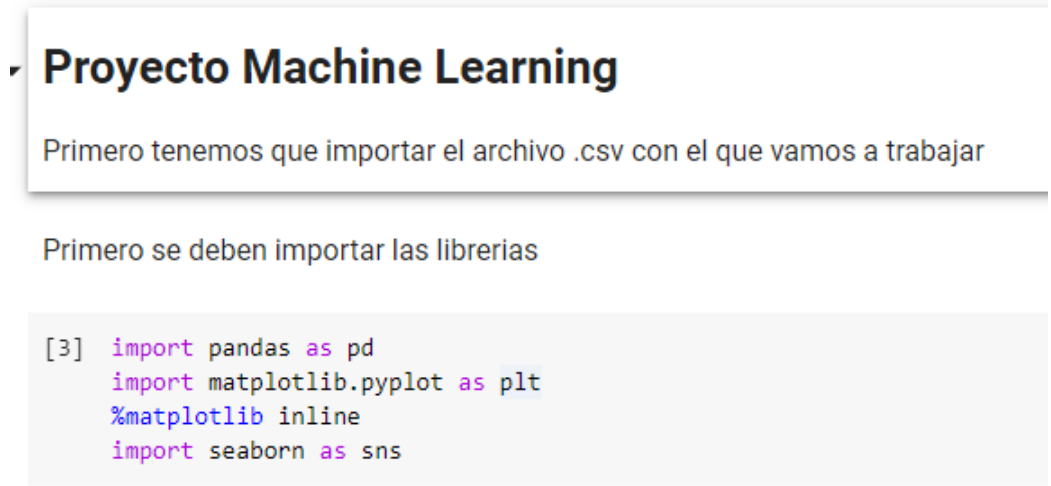

Figura 1: Importación de librerías

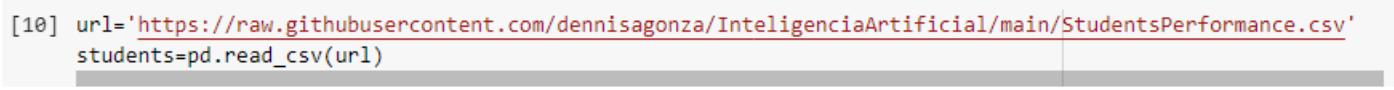

Figura 2: Importación del archivo

Para empezar podemos ver las primeras filas de la tabla para poder ver que datos se estan presentando en este archivo

```
[7]  students.head()
```

	gender	race/ethnicity	parental level of education	lunch	test preparation course	math score	reading score	writing score
0	female	group B	bachelor's degree	standard	none	72	72	74
1	female	group C	some college	standard	completed	69	90	88
2	female	group B	master's degree	standard	none	90	95	93
3	male	group A	associate's degree	free/reduced	none	47	57	44
4	male	group C	some college	standard	none	76	78	75

Figura 3: Visualización de las primeras filas de la tabla

A partir de la gráfica anterior, filtramos los tres mejores puntajes los cuales podemos observar que dos son femeninos y uno masculino, los cuales pertenecen al grupo étnico E, las tres personas sacaron en todas las materias un puntaje máximo de 100 y solo el hombre realizo un curso de preparación para el mismo.

Creamos una variable donde se sumarian los tres puntajes por estudiante y asi poder analizar mucho mas facil el rendimiento de estos en general, tambien filtramos los estudiantes que obtuvieron el mayor puntaje total e imprimos todos sus datos para su comparación

```
[ ]  PuntajeTotal=students['math score']+students['writing score']+students['reading score']
     print(PuntajeTotal[PuntajeTotal==300])
```

```
458    300
916    300
962    300
dtype: int64
```

Figura 4: Visualización de datos filtrados para su correcta comparación y análisis

```
[ ]  students[458:459]
```

	gender	race/ethnicity	parental level of education	lunch	test preparation course	math score	reading score	writing sco
458	female	group E	bachelor's degree	standard	none	100	100	1

Figura 5: Visualización de datos filtrados

```
[ ]  students[916:917]
```

	gender	race/ethnicity	parental level of education	lunch	test preparation course	math score	reading score	writing score
916	male	group E	bachelor's degree	standard	completed	100	100	100

Figura 6: Visualización de datos filtrados

```
[ ] students[962:963]
```

Agregar celda de texto

	race/ethnicity	parental level of education	lunch	test preparation course	math score	reading score	writing score
962	female	group E	associate's degree	standard	none	100	100	100

Figura 7: Visualización de datos filtrados

A continuación, podremos observar en las siguientes gráficas la descripción de cada información de la columna de *Puntajes*.

En los cuales se puede observar que en todas las materias se obtuvo una calificación de 100 como puntaje máximo, mientras que la media varia entre 66 a 69. Y, como puntaje mínimo se obtiene una variación de las calificaciónes entre 0 a 17, el cual matemáticas obtuvo el menor puntaje.

El promedio de las calificaciones de los examenes en la materia de matemáticas es de 66, en el cual la calificación mínima de 17 y la máxima es de 100.

Tambien se describio la informacion de cada columna de Puntajes, para obtener informacion especifica como lo fue el mayor y menor puntaje de cada examen y el promedio de estos.

```
[ ] students['math score'].describe()

count    1000.00000
mean       66.08900
std        15.16308
min         0.00000
25%        57.00000
50%        66.00000
75%        77.00000
max       100.00000
Name: math score, dtype: float64
```

Figura 8: Descripción de la información de datos de los puntajes del área de matemáticas

El promedio de las calificaciones de los examenes en la materia de lectura es de 69, en el cual la calificación mínima de 17 y la máxima es de 100

```
students['reading score'].describe()

count    1000.000000
mean       69.169000
std        14.600192
min        17.000000
25%        59.000000
50%        70.000000
75%        79.000000
max       100.000000
Name: reading score, dtype: float64
```

Figura 9: Descripción de la información de datos de los puntajes del área de lectura

El promedio de las calificaciones de los examenes en la materia de escritura es de 68, en el cual la calificación mínima es de 10 y la máxima es de 100

```
[ ] students['writing score'].describe()

count    1000.000000
mean       68.054000
std        15.195657
min        10.000000
25%        57.750000
50%        69.000000
75%        79.000000
max       100.000000
Name: writing score, dtype: float64
```

Figura 10: Descripción de la información de datos de los puntajes del área de escritura

En esta gráfica se puede interpretar que el genero femenino obtuvo un mejor puntaje en todas las materias que el genero masculino.

¿Que genero obtuvo el mejor puntaje?

```
sns.catplot(x='gender', y=PuntajeTotal, kind='bar', data=students)
plt.title("Puntaje Total respecto al genero")
plt.ylabel("Puntaje Total")
plt.xlabel("Genero")

Text(0.5, 6.79999999999999, 'Genero')
```

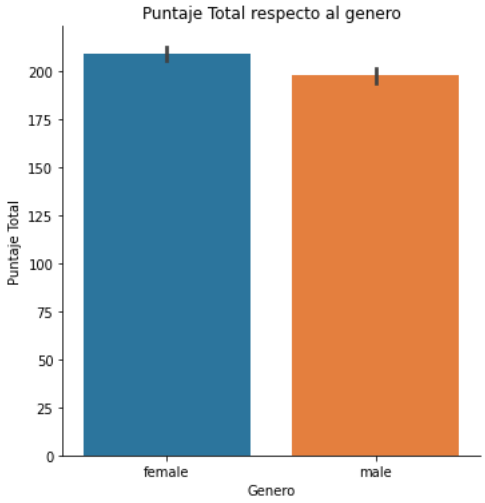

Figura 11: Gráfica de puntaje en el examen respecto a cada genero

Para la siguiente gráfica se tomo el puntaje de cada examen respecto a si habian completado el test de preparación del curso. Como se puede evidenciar, si realizan un test de preparación no significa que van obtener los mejores puntajes, pero si se evidencia algún conocimiento previo puesto que no tienen tan bajos puntajes respecto a las personas que no realizaron el test.

Por otro lado, se demuestra que evidentemente son muchos más estudiantes que obtuvieron los mejores puntajes si realizaron un test de preparación anteriormente, en cambio las personas que no lo realizaron estan sobre la media total de los puntajes.

¿Hay relacion entre los puntajes obtenidos y que completaran un test de preparacion para el curso?

```
[ ] sns.swarmplot(x=students['test preparation course'], y=PuntajeTotal)
    plt.title("Puntaje total en relación a un Test de preparación al curso")
    plt.ylabel("Puntaje Total")
    plt.xlabel("Test de preparación")
```

```
/usr/local/lib/python3.7/dist-packages/seaborn/categorical.py:1296: UserWarning: 20.1% of
  warnings.warn(msg, UserWarning)
Text(0.5, 0, 'Test de preparación')
```

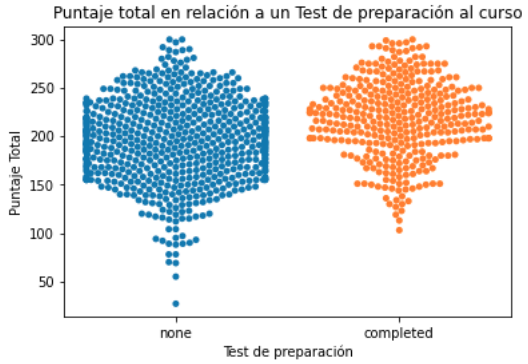

Figura 12: Gráfica de puntaje en el examen respecto a el examen de preparación del curso

En la siguiente gráfica se puede visualizar que grupos étnicos tuvieron un mejor puntaje en todas las áreas de aprendizaje, como: matemáticas, lectura y escritura. En las cuales, podemos concluir que el grupo étnico E sobresale en las tres materias con puntajes altos, mientras que el grupo étnico A tiene los menores puntajes en todas las materias.

¿Que grupos etnicos tuvieron mejor puntaje?

```
[ ] sns.catplot(x='race/ethnicity', y=PuntajeTotal, kind='bar', data=students)
    plt.title("Puntaje Total respecto a su Raza/Etnia")
    plt.ylabel("Puntaje Total")
    plt.xlabel("Raza/Etnia")
```

```
Text(0.5, 6.79999999999999, 'Raza/Etnia')
```

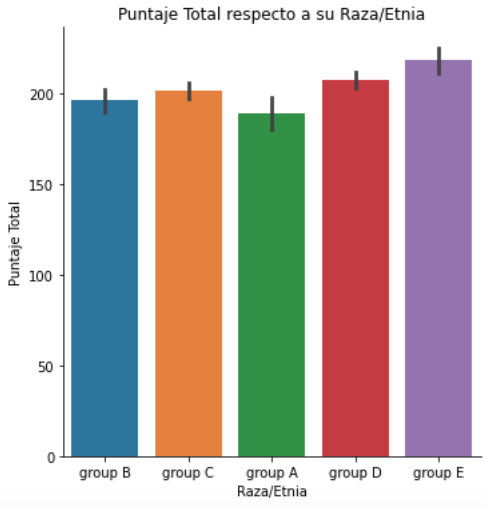

Figura 13: Gráfica de puntaje en cada grupo étnico

Para la gráfica a continuación se realizó un histograma por cada área: matemáticas. lectura y escritura, el cual se utiliza para identificar e interpretar cantidades de datos. Teniendo en cuenta lo anterior podemos concluir que en todos los

histogramas visualizados en el presente documento la dispersión de datos va desde la nota mínima de 0 y la nota mas alta de 100. En donde, los puntajes de cada área se centran o tienen un pico entre el puntaje de 60 y 80.

Generamos un histograma para cada área

```
pd.DataFrame.hist(students, layout=(7,2), figsize=(15,15), bins=50)
```

```
array([[<matplotlib.axes._subplots.AxesSubplot object at 0x7f5f0fa5a950>,
        <matplotlib.axes._subplots.AxesSubplot object at 0x7f5f0f8ff750>],
       [<matplotlib.axes._subplots.AxesSubplot object at 0x7f5f0f8b2cd0>,
        <matplotlib.axes._subplots.AxesSubplot object at 0x7f5f0f873290>],
       [<matplotlib.axes._subplots.AxesSubplot object at 0x7f5f0f82a810>,
        <matplotlib.axes._subplots.AxesSubplot object at 0x7f5f0f85cd90>],
       [<matplotlib.axes._subplots.AxesSubplot object at 0x7f5f0f81f3d0>,
        <matplotlib.axes._subplots.AxesSubplot object at 0x7f5f0f7d4890>],
       [<matplotlib.axes._subplots.AxesSubplot object at 0x7f5f0f7d48d0>,
        <matplotlib.axes._subplots.AxesSubplot object at 0x7f5f0f789f50>],
       [<matplotlib.axes._subplots.AxesSubplot object at 0x7f5f0f703990>,
        <matplotlib.axes._subplots.AxesSubplot object at 0x7f5f0f6b9f10>],
       [<matplotlib.axes._subplots.AxesSubplot object at 0x7f5f0f67c4d0>,
        <matplotlib.axes._subplots.AxesSubplot object at 0x7f5f0f631a50>]],
      dtype=object)
```

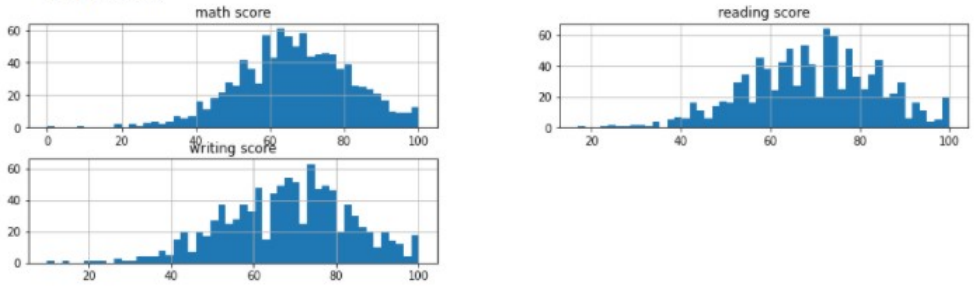

Figura 14: Histograma de cada área

En este paso, establecemos como etiqueta al área de matemáticas. La cual nos establecera una guía u orden en los datos proximos al mismo. Y, también se establece que no se vuelva a guardar los datos actualizados en la misma variable.

```
[47] stu_labels=students['math score']
```

```
[48] stu_labels
```

```
0      72
1      69
2      90
3      47
4      76
      ..
995    88
996    62
997    59
998    68
999    77
Name: math score, Length: 1000, dtype: int64
```

```
[49] students.drop('math score',axis=1,inplace=True)
```

Figura 15: Histograma de cada área

En esta figura se puede observar que no se guardaron los datos actualizados en la variable establecida anteriormente de matemáticas.

```
students
```

	race/ethnicity	parental level of education	lunch	test preparation course	reading score	writing score
0	group B	bachelor's degree	standard	none	72	74
1	group C	some college	standard	completed	90	88
2	group B	master's degree	standard	none	95	93
3	group A	associate's degree	free/reduced	none	57	44
4	group C	some college	standard	none	78	75
...	...	...	...	...	...	...
995	group E	master's degree	standard	completed	99	95
996	group C	high school	free/reduced	none	55	55
997	group C	high school	free/reduced	completed	71	65
998	group D	some college	standard	completed	78	77
999	group D	some college	free/reduced	none	86	86

1000 rows × 6 columns

Figura 16: Histograma de cada área

En la siguiente imagen, se combino los puntajes de las áreas de escritura y lectura en las cuales se muestra las características cuantitativas de las mismas.

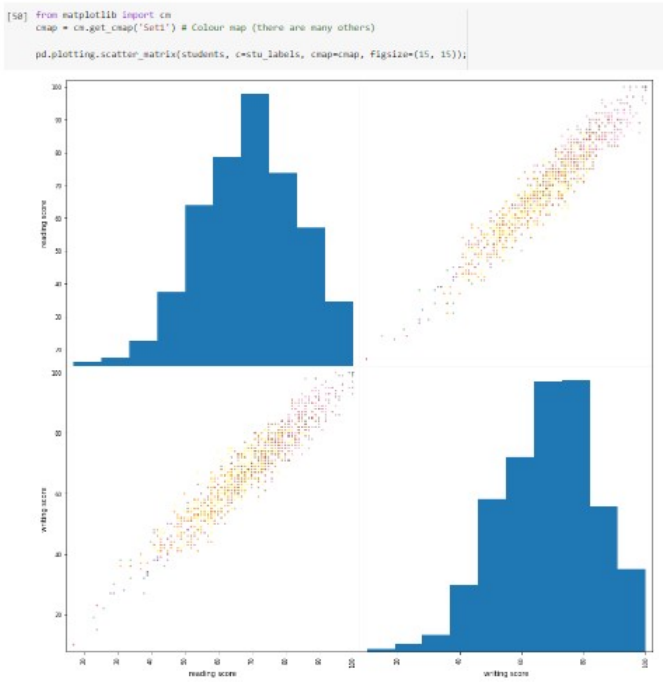

Figura 17: Combinación de puntajes del área de lectura y escritura

Creamos una variable en la cual se le guardará la operación en la cual se cambia el valor de los datos de genero por números, los cuales *femenino* es igual a cero, y *masculino* es igual a 1. Y, también se eliminan las columnas siguientes a las del genero con el fin de tener un mejor manejo de los datos.

```
[72] stu_labels = students['gender'].replace(['female','male'],[0,1])
     students.drop('gender',axis=1,inplace=True)
     students.drop('race/ethnicity',axis=1,inplace=True)
     students.drop('parental level of education',axis=1,inplace=True)
     students.drop('lunch',axis=1,inplace=True)
     students.drop('test preparation course',axis=1,inplace=True)

[73] X=np.array(students)
     y=np.array(stu_labels)
```

Figura 18: Creación de una nueva variable con sus respectivos ajustes

A continuación se viasualiza la salida de la variable y ajustes creados anteriormente.

```
stu_labels

0      0
1      0
2      0
3      1
4      1
      ..
995    0
996    1
997    0
998    0
999    0
Name: gender, Length: 1000, dtype: int64
```

Figura 19: Salida de la variable creada

Se crean dos variables, las cuales se les asigna ciertos valores, y creamos una gráfica en la cual se combinan los géneros, por cada área nombrada anteriormente, en los cuales se muestran sus características cuantitativas.

Figura 20: Gráfica de género por cada área

3. Conclusiones

- Al filtrar por los tres mejores puntajes de todos los estudiantes, se encontró a dos mujeres y dos hombres. Los cuales pertenecen al grupo étnico E, y solo el hombre realizó el curso de preparación para el ingreso al mismo.
- Los tres mejores estudiantes, obtuvieron en todas las asignaturas un puntaje máximo de 100 puntos.
- En las tres materias se obtuvo un puntaje máximo de 100 puntos, mientras que el puntaje mínimo se encuentra en en un rango de 0 a 17 puntos.
- En relación con todos los puntajes filtrado por generos, se encontró que el genero femenino tuvo un mayor puntaje en todas las materias que el genero masculino.
- El test de preparación se realiza para obtener conocimientos previos de las materias y no garantiza que saquen los mayores puntajes.
- Los puntajes filtrados por cada grupo étnico demuestran que el grupo E tiene los mayores puntajes, mientras que el grupo A los menores.
- Los mayores puntajes de los estudiantes en las áreas de matemáticas, lectura y escritura se encuentran entre 60 y 80, con una mayor asistencia de puntajes en el área de las matemáticas.

Referencias

[1] J.Seshapanpu “Students Performance in Exams” Kaggle, url: https://www.kaggle.com/spscientist/students-performance-in-exams, Noviembre 2018.

[2] Repositorio en Git https://github.com/dennisagonza/InteligenciaArtificial