Análisis de los Hábitos Estudiantiles y su Impacto en el Rendimiento Académico

Oscar Jared Cruz Gozalez

202478335 Prof.Jaime Alejandro Romero Sierra

Desarrollar un análisis de datos sobre hábitos estudiantiles para poder predecir si un estudiante podrá aprobar sus exámenes. Este sistema ayudará a los estudiantes a identificar áreas de mejora en sus hábitos académicos.

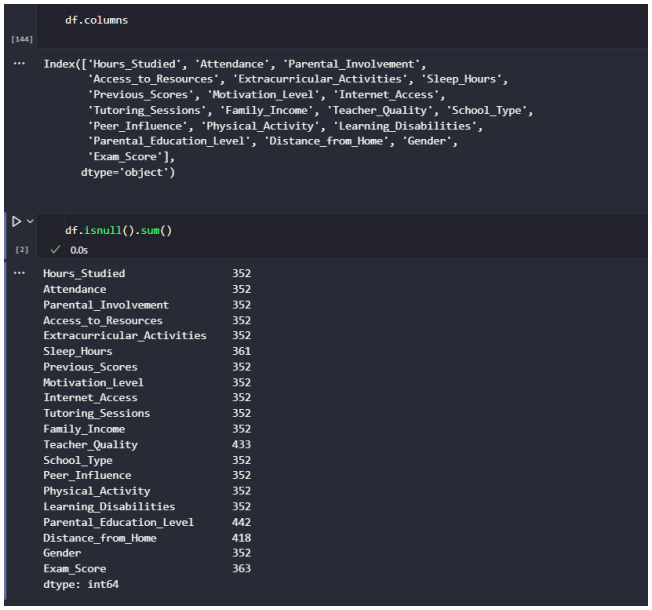
Justificación y contexto:
Muchos estudiantes no conocen el impacto de sus hábitos en el rendimiento académico. Este proyecto analiza cómo factores como las horas de estudio, la asistencia y el apoyo familiar influyen en los resultados escolares. Al proporcionar un diagnóstico predictivo, se busca fomentar mejores hábitos, reducir el estrés y mejorar las tasas de aprobación.

Fuentes de datos:

- La base de datos contiene 7053 registros y 20 columnas
- Las variables incluyen:
    - Hábitos estudiantiles: Horas de Estudio, Asistencia, Horas de Sueño, Actividades Extracurriculares.
    - Factores familiares: Involucramiento de los Padres, Ingreso Familiar, Nivel de Educación de Padres.
    - Factores escolares: Calidad del Maestro, Tipo de Escuela, Sesiones de Tutorias.
    - Factores individuales: Nivel de Motivacion, Influencias de Compañeros, Actividad Fisica, Discapacidad de Aprendizaje.
    - Resultados académicos: Puntaje de Examen, Puntajes Anteriores.

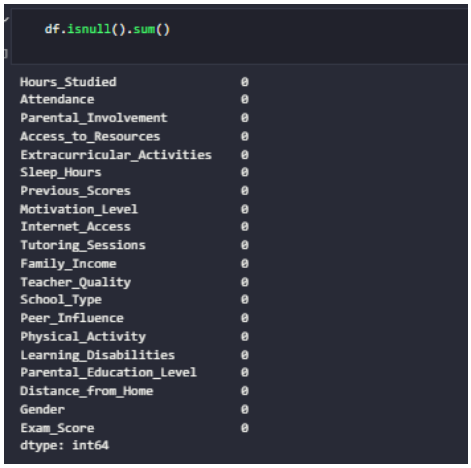
- Proceso de limpieza de datos:
    - Identificación y manejo de valores ausentes.
      Se realizó un análisis para detectar valores faltantes (NaN) en cada columna usando herramientas de conteo como isnull().sum().
    - Se encontraron valores ausentes y se asignó una nueva categoría llamada "Desconocido" para preservar la integridad de los datos.

```
    "Attendance",
    "Sleep_Hours",
    "Previous_Scores",
    "Tutoring_Sessions",
    "Physical_Activity",
    "Exam_Score","Hours_Studied",]
[45]

for columna in columnas_a_limpiar:
    df[columna] = df[columna].replace(" ", 0)
    df[columna] = df[columna].fillna(0)
[46]

colText = [
    "Parental_Involvement",
    "Access_to_Resources",
    "Extracurricular_Activities",
    "Motivation_Level",
    "Internet_Access",
    "Teacher_Quality",
    "School_Type",
    "Peer_Influence",
    "Learning_Disabilities",
    "Parental_Education_Level",
    "Gender","Distance_from_Home","Family_Income"
]
[47]

for columna in colText:
    df[columna] = df[columna].fillna("Desconocido")
    df[columna] = df[columna].str.strip()
[48]
```



```
    df.columns
[144]

... Index(['Hours_Studied', 'Attendance', 'Parental_Involvement',
       'Access_to_Resources', 'Extracurricular_Activities', 'Sleep_Hours',
       'Previous_Scores', 'Motivation_Level', 'Internet_Access',
       'Tutoring_Sessions', 'Family_Income', 'Teacher_Quality', 'School_Type',
       'Peer_Influence', 'Physical_Activity', 'Learning_Disabilities',
       'Parental_Education_Level', 'Distance_from_Home', 'Gender',
       'Exam_Score'],
      dtype='object')
```

```
    df.isnull().sum()
[2] ✓ 0.0s

... Hours_Studied                352
    Attendance                   352
    Parental_Involvement         352
    Access_to_Resources          352
    Extracurricular_Activities   352
    Sleep_Hours                  361
    Previous_Scores              352
    Motivation_Level             352
    Internet_Access              352
    Tutoring_Sessions            352
    Family_Income                352
    Teacher_Quality              433
    School_Type                  352
    Peer_Influence               352
    Physical_Activity            352
    Learning_Disabilities        352
    Parental_Education_Level     442
    Distance_from_Home           418
    Gender                       352
    Exam_Score                   363
    dtype: int64
```



```
    df.isnull().sum()

Hours_Studied                0
Attendance                   0
Parental_Involvement         0
Access_to_Resources          0
Extracurricular_Activities   0
Sleep_Hours                  0
Previous_Scores              0
Motivation_Level             0
Internet_Access              0
Tutoring_Sessions            0
Family_Income                0
Teacher_Quality              0
School_Type                  0
Peer_Influence               0
Physical_Activity            0
Learning_Disabilities        0
Parental_Education_Level     0
Distance_from_Home           0
Gender                       0
Exam_Score                   0
dtype: int64
```

1. Descripción General de los Datos:
    1. Registros: 7053.
    2. Columnas: 20.
    3. Clasificación:
    - Variables numéricas: 'Horas_Estudio', 'Asistencia','Horas_Sueño', 'Puntajes_Anteriores', 'Actividad_Fisica', 'Puntaje_Examen'
    - Variables categóricas:'Acceso_a_Internet', 'Calidad_Maestro', 'Tipo_Escuela','Influencias_Compañeros', 'Actividad_Fisica','Discapacidad_de_Aprendizaje', 'Nivel_Educacion_Padres', 'Genero'

- Transformación de Variables Categóricas
    - Las variables categóricas se convirtieron al tipo category para optimizar el uso de memoria y facilitar el análisis.

```
columnas_numericas = [
    'Horas_Estudio', 'Asistencia','Horas_Sueño', 'Puntajes_Anteriores',
    'Actividad_Fisica', 'Puntaje_Examen'
]
```

```
columnas_categoricas = [
    'Involucramiento_Padres', 'Acceso_a_Recursos', 'Actividades_Extracurriculares',
    'Acceso_a_Internet', 'Calidad_Maestro', 'Tipo_Escuela',
    'Influencias_Compañeros', 'Actividad_Fisica', 'Discapacidad_de_Aprendizaje',
    'Nivel_Educacion_Padres', 'Genero'
]
```

- Normalización y Escalado
    - Variables escaladas: Las variables numéricas con diferentes rangos fueron normalizadas utilizando Min-Max para garantizar la comparabilidad en los análisis

```
columnas_numericas = [
    'Horas_Estudio', 'Asistencia','Horas_Sueño', 'Puntajes_Anteriores',
    'Actividad_Fisica', 'Puntaje_Examen'
]

resumen_numerico = df[columnas_numericas].describe().transpose()
print("Resumen de variables numéricas:")
print(resumen_numerico[['count', 'mean', 'std', 'min', '25%', '50%', '75%', 'max']])
```

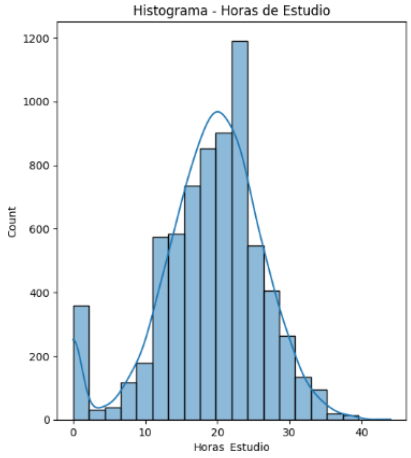
```
Resumen de variables numéricas:
                    count       mean        std  min   25%   50%   75%  \
Horas_Estudio      7046.0  18.975163   7.315215  0.0  15.0  20.0  24.0
Asistencia         7045.0  75.950887  20.754052  0.0  68.0  79.0  90.0
Horas_Sueño        7053.0   6.672763   2.108808  0.0   6.0   7.0   8.0
Puntajes_Anteriores 7045.0  71.318808  21.563292  0.0  61.0  74.0  87.0
Actividad_Fisica   7049.0   2.822812   1.197785  0.0   2.0   3.0   4.0
Puntaje_Examen     7053.0  63.763363  15.330357  0.0  64.0  67.0  69.0

                     max
Horas_Estudio       44.0
Asistencia         100.0
Horas_Sueño         10.0
Puntajes_Anteriores 100.0
Actividad_Fisica      6.0
Puntaje_Examen      101.0
```

```
columnas_categoricas = [
    'Involucramiento_Padres', 'Acceso_a_Recursos', 'Actividades_Extracurriculares',
    'Acceso_a_Internet', 'Calidad_Maestro', 'Tipo_Escuela',
    'Influencias_Compañeros', 'Actividad_Fisica', 'Discapacidad_de_Aprendizaje',
    'Nivel_Educacion_Padres', 'Genero'
]

print("Frecuencia de variables categóricas:")
for columna in columnas_categoricas:
    print(f"\nFrecuencia para {columna}:")
    print(df[columna].value_counts())
```

```
Frecuencia de variables categóricas:

Frecuencia para Involucramiento_Padres:
Involucramiento_Padres
Media          3419
Alta           1925
Baja           1348
Desconocido     352
invalid           9
Name: count, dtype: int64

Frecuencia para Acceso_a_Recursos:
Acceso_a_Recursos
Media          3359
Alta           2001
Baja           1337
Desconocido     352
invalid           4
Name: count, dtype: int64

Frecuencia para Actividades_Extracurriculares:
Actividades_Extracurriculares
Si             3984
No             2712
Desconocido     352
...
Female         2818
Desconocido     352
invalid           8
Name: count, dtype: int64
```

2. Relación entre Variables Categóricas y Numéricas: Visualización y Distribución de Variables Individuales:

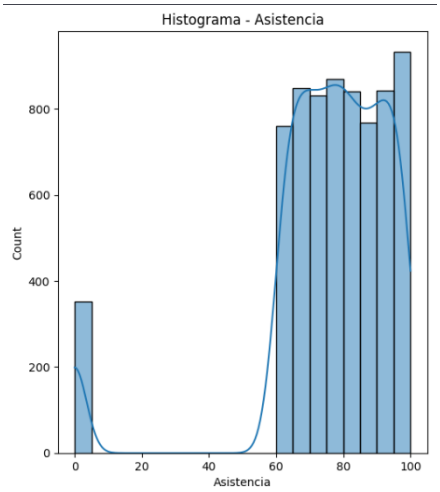
- Variables numéricas:
- Horas de Estudio



Eje X: Representa la cantidad de horas que los estudiantes dedican al estudio

Eje Y: Cuenta o frecuencia de estudiantes en cada rango de horas de estudio.

La mayoría de los estudiantes estudian entre 15 y 25 horas por semana, con un pico cercano a las 20 horas. La distribución disminuye hacia la derecha, indicando que hay pocos estudiantes que dedican más de 30 horas al estudio



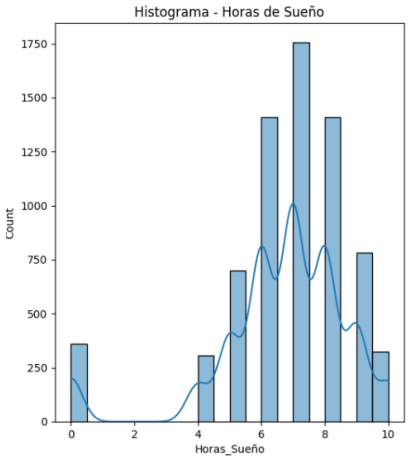
- Asistencia

Eje X: Representa el porcentaje de asistencia de los estudiantes de 0% a 100%.

Eje Y: Frecuencia de estudiantes para cada rango de asistencia.

La mayoría de los estudiantes tienen asistencias entre el 80% y el 100%, muy pocos estudiantes tienen asistencias por debajo del 50%, lo que sugiere que la falta de asistencia extrema no es común en la muestra

- Horas de Sueño



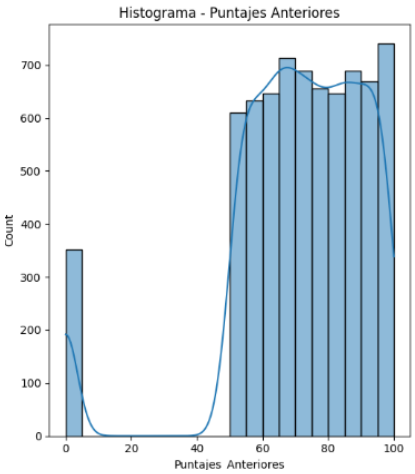
Eje X: Número de horas de sueño diarias (de 0 a 10 horas).

Eje Y: Frecuencia de estudiantes en cada rango de horas de sueño.

Observaciones:

La distribución parece bimodal, con picos alrededor de 6 horas y 8 horas de sueño.

Algunos estudiantes duermen muy poco (0-2 horas), lo que podría afectar su rendimiento.



- Puntajes Anteriores

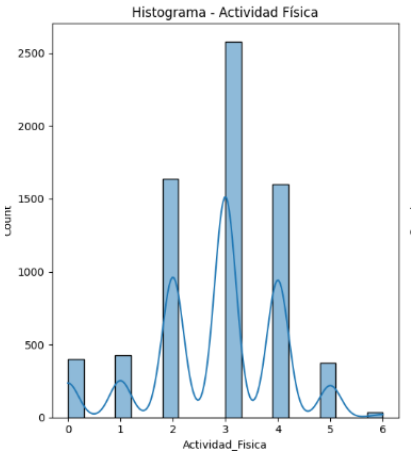
Eje X: Puntajes obtenidos en evaluaciones anteriores (de 0 a 100).

Eje Y: Frecuencia de estudiantes para cada rango de puntajes.

La mayoría de los puntajes se encuentran entre 40 y 90.

Existe un pequeño grupo de estudiantes con puntajes cercanos a 0, lo que podría indicar bajo desempeño o dificultades previas

- Actividad Fisica



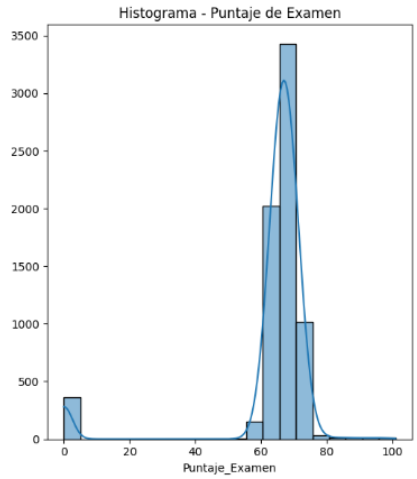
Eje X: Nivel de actividad física semanal (puede ser categórico o en niveles).

Eje Y: Frecuencia de estudiantes en cada nivel de actividad física.

Observaciones:

Hay un pico en el nivel de actividad 3, indicando que la mayoría de los estudiantes tienen un nivel moderado de actividad.

Los niveles extremos (1 y 6) tienen frecuencias bajas.



- Puntaje de Examen

Eje X: Puntajes obtenidos en exámenes actuales (de 0 a 100).

Eje Y: Frecuencia de estudiantes para cada rango de puntajes.

Observaciones:

La distribución está centrada alrededor de puntajes entre 60 y 80.

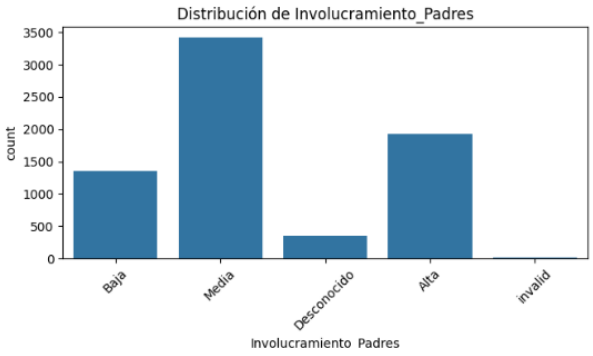
Muy pocos estudiantes obtienen puntajes cercanos a 0 o 100.

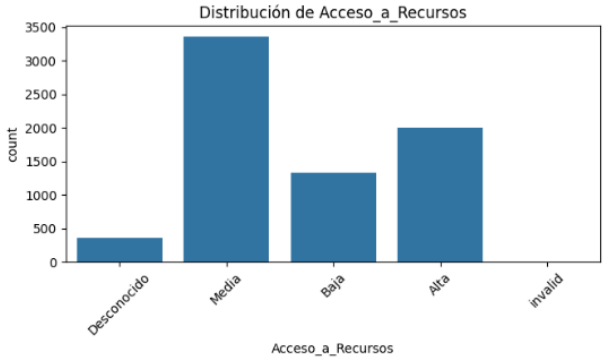
- Variables categóricas: Gráficos de barras para analizar frecuencias.

- Distribución de Involucramiento_Padres

La mayoría de los estudiantes tienen un involucramiento parental medio, con más de 3000 observaciones en esta categoría.

El nivel de involucramiento alto es el siguiente más común, mientras que los niveles bajos y desconocidos son menos frecuentes y los datos clasificados como inválidos son casi inexistentes, esto indica que la participación de los padres en la educación de sus hijos suele ser moderada o alta, lo cual podría influir positivamente en el rendimiento académico de los estudiantes.





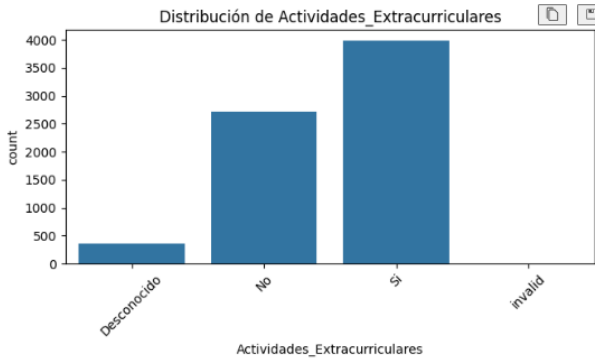
- Distribución de Acceso_a_Recursos

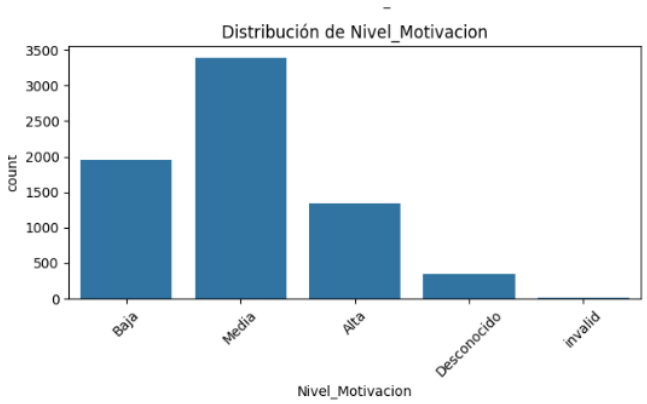
El acceso medio a recursos es el más común, con más de 3500 observaciones. Los estudiantes con acceso alto también representan un grupo considerable, mientras que el acceso bajo es menos frecuente. Las categorías desconocido e inválido tienen muy pocos registros.
Esto refleja que, aunque una proporción significativa de estudiantes tiene un nivel moderado o alto de acceso a recursos, todavía existe una minoría con acceso limitado, lo que podría dificultar su aprendizaje.

- Distribución de Actividades_Extracurriculares

Más de 4000 estudiantes participan en actividades extracurriculares (categoría "Sí"). Alrededor de 3000 estudiantes no participan en estas actividades. Las categorías desconocido e inválido son mínimas.

La participación en actividades extracurriculares es alta, lo cual puede fomentar habilidades adicionales en los estudiantes. Sin embargo, la considerable proporción que no participa podría deberse a factores como limitaciones de tiempo o falta de recursos.





- Distribución de Nivel_Motivacion

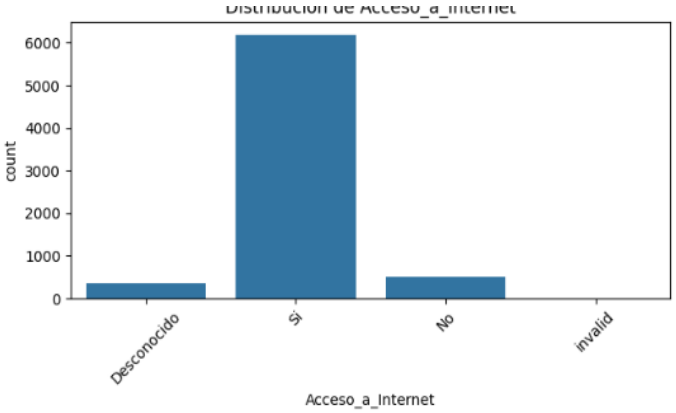
El nivel de motivación medio es predominante, con más de 3500 estudiantes. La motivación baja es el segundo grupo más común, mientras que la motivación alta es menos frecuente.

Aunque la mayoría de los estudiantes tienen una motivación moderada, la presencia de un grupo significativo con baja motivación podría impactar negativamente en su desempeño académico.

- Distribución de Acceso_a_Internet

La mayoría de los estudiantes tiene acceso a Internet, con aproximadamente 6000 observaciones en la categoría "Sí". Una minoría considerable no cuenta con acceso a Internet.

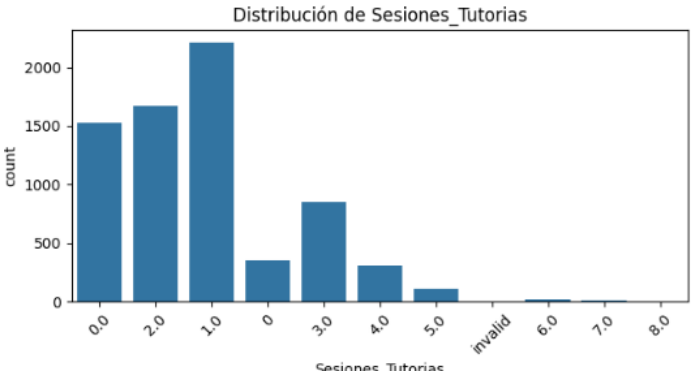
El acceso a Internet es casi universal entre los estudiantes, lo que les brinda oportunidades para utilizar herramientas digitales en su aprendizaje. Sin embargo, es necesario considerar soluciones para aquellos que carecen de acceso.

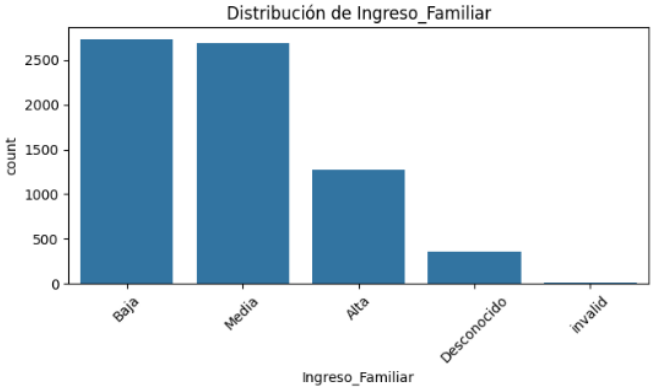


- Distribución de Sesiones_Tutorias

La mayoría de los estudiantes asisten a 2 o 3 sesiones de tutoría, con un pico en 2 sesiones. El número de estudiantes que asisten a más de 5 sesiones disminuye significativamente.

La tutoría es una práctica extendida entre los estudiantes, pero su frecuencia podría ser insuficiente para aquellos con mayores necesidades de apoyo.





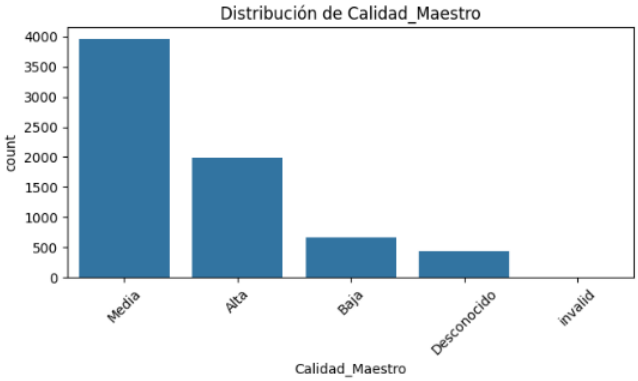
- Distribución de Ingreso_Familiar

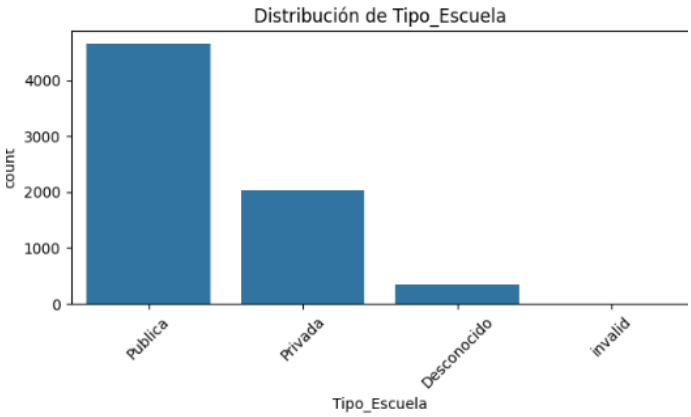
Las categorías de ingresos bajos y medios son las más comunes, cada una con más de 2500 observaciones. La categoría de ingresos altos tiene una representación mucho menor. Una gran proporción de los estudiantes proviene de familias con ingresos bajos o medios, lo cual puede limitar el acceso a recursos educativos y oportunidades adicionales. Esto subraya la importancia de iniciativas de apoyo financiero y educativo.

- Distribución de Calidad_Maestro

La percepción de calidad media de los maestros es la más común, seguida por la calidad alta. Las percepciones de calidad baja son menos frecuentes.

Esto sugiere que la mayoría de los estudiantes consideran a sus maestros como moderadamente competentes o excelentes. Sin embargo, las percepciones de baja calidad podrían impactar negativamente en el aprendizaje y deben ser atendidas.





- Distribución de Tipo_Escuela

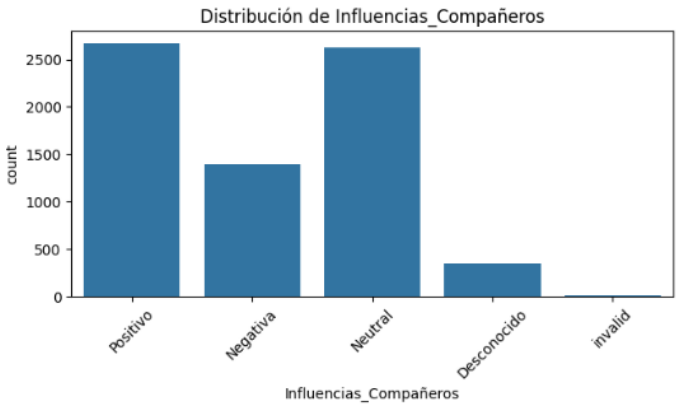
La mayoría de los estudiantes asisten a escuelas públicas, con más de 4000 observaciones en esta categoría. Un número menor asiste a escuelas privadas.

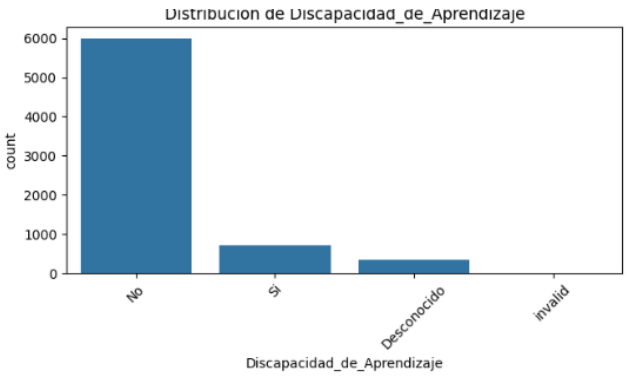
Esto refleja la predominancia del sistema público en la educación de los estudiantes, destacando la necesidad de enfoques específicos para atender las limitaciones y oportunidades de este tipo de institución.

- Distribución de Influencias_Compañeros

La mayoría de las observaciones se clasifican como "Positivo" y "Neutral", con más de 2500 registros en cada categoría. "Negativa" tiene un número considerable, aunque significativamente menor.

Esto resalta que las influencias de los compañeros suelen percibirse como positivas o neutrales, lo que puede reflejar un entorno social generalmente favorable en el ámbito estudiantil.





- Distribución de Discapacidad_de_Aprendizaje

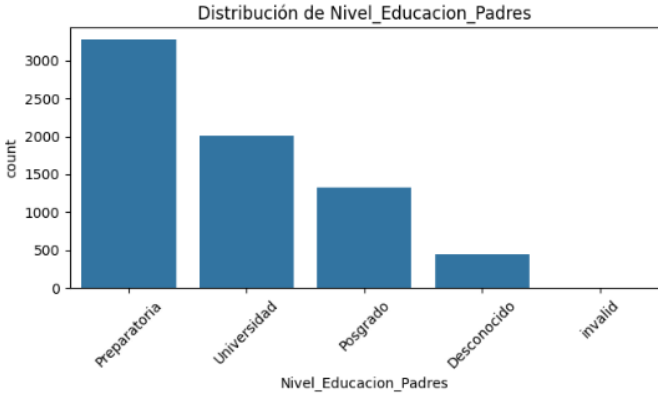
La mayoría de los datos están etiquetados como "No", con más de 5000 observaciones, indicando que la gran mayoría de los estudiantes no presentan discapacidades de aprendizaje. Una proporción menor está clasificada como "Sí".

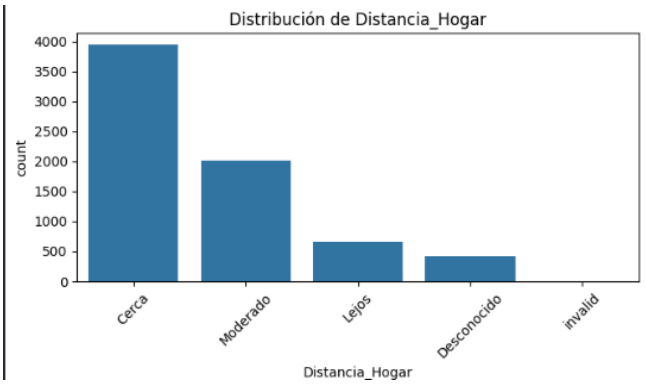
Esto sugiere que la inclusión de estudiantes con discapacidades de aprendizaje puede ser un área de mejora, o que existen desafíos en su identificación y registro.

- Distribución de Nivel_Educacion_Padres

El nivel educativo más común es "Preparatoria", con más de 3000 registros, seguido de "Universidad" y "Posgrado" con menos representación.

Esto refleja que la mayoría de los padres tienen una formación de nivel medio-superior, lo que podría influir en las expectativas y el apoyo académico que los estudiantes reciben en casa.





- Distribución de Distancia_Hogar

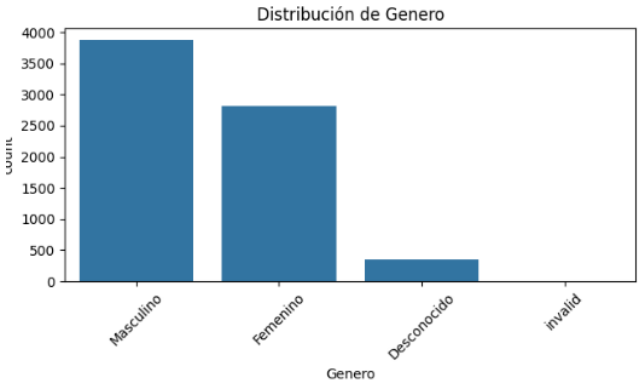
La mayoría de los estudiantes viven cerca de su lugar de estudio, con más de 3500 observaciones, mientras que la categoria "Moderado" tiene un número considerablemente menor. Y "Lejos" tiene una representación mínima, comparado a los demas.

Esto sugiere que la cercanía al hogar es un factor importante en la educación, posiblemente relacionado con la accesibilidad y la elección de las instituciones educativas.

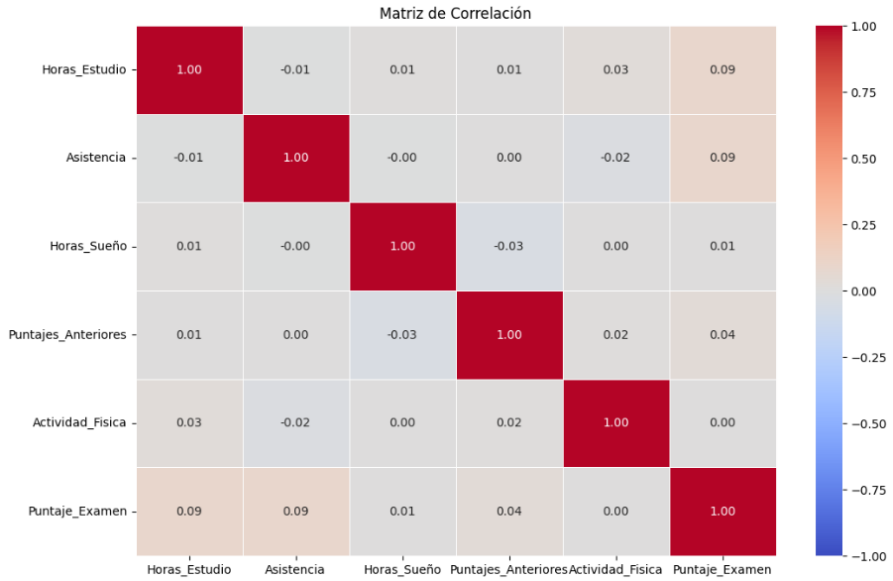
- Distribución de Género

La categoría "Masculino" es predominante, con aproximadamente 4000 registros, mientras que “Femenino” cuenta con menos observaciones, aunque sigue siendo significativa.

Esto indica una ligera disparidad de género en los datos, que podría deberse a factores contextuales, culturales o de muestreo.



3. Correlación entre Variables



La gráfica muestra la matriz de correlación entre diferentes variables relacionadas con el desempeño estudiantil. Cada celda representa el coeficiente de correlación entre dos variables, que varía entre -1 y 1. Los colores indican la intensidad y dirección de la correlación, desde azul (negativa) hasta rojo oscuro (positiva).

- Horas_Estudio

Tiene una ligera correlación positiva con "Puntaje_Examen" (0.09), Esto quiere decir que dedicar más tiempo al estudio podría estar relacionado con un mejor desempeño en los exámenes, aunque el efecto es pequeño,las correlaciones cercanas a 0 con el resto de las variables, lo que indica una relación muy débil o inexistente.

- Asistencia

Muestra una correlación positiva baja con "Puntaje_Examen" (0.09), esto implica que una mejor asistencia podría influir levemente en un mejor rendimiento académico, pero prácticamente no tiene relación con "Horas_Estudio", "Horas_Sueño", ni otras variables.

- Horas_Sueño

No presenta correlaciones significativas con ninguna variable, esto indica que la cantidad de horas de sueño no parece influir directamente en los resultados de exámenes, estudio, o actividad física en este conjunto de datos.

- Puntajes_Anteriores

Tiene una correlación muy baja con "Puntaje_Examen" (0.04), esto sugiere que los puntajes previos tienen poca influencia en el desempeño en exámenes actuales y carece de correlación relevante con el resto de las variables.

- Actividad_Física

No muestra correlaciones significativas con ninguna variable, esto implica que la cantidad de actividad física realizada no tiene un impacto notable en el estudio, sueño, asistencia, ni puntajes.

- Puntaje_Examen

Tiene correlaciones positivas bajas con "Horas_Estudio" (0.09) y "Asistencia" (0.09),esto indica que tanto estudiar como asistir regularmente contribuyen marginalmente al desempeño en exámenes, pero no tiene correlación relevante con "Horas_Sueño", "Puntajes_Anteriores", ni "Actividad_Física".

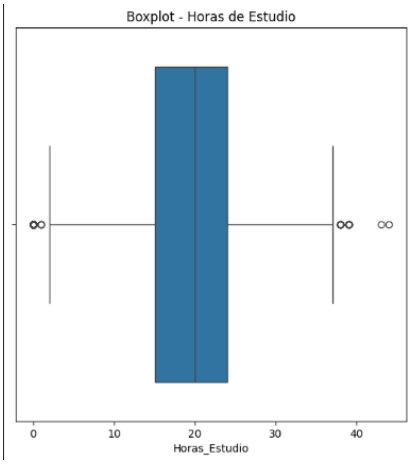
- Conclusión

La matriz muestra que no hay correlaciones fuertes entre las variables, pero se identifican relaciones débiles entre el desempeño en exámenes y factores como horas de estudio y Asistencia . Esto sugiere que, aunque estos factores tienen cierta influencia, existen otros elementos que podrían ser más determinantes en el rendimiento académico.

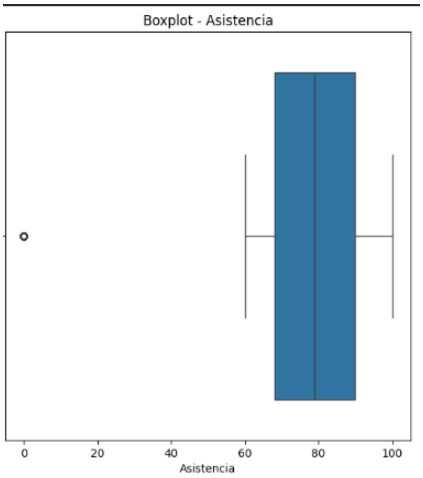
4. Análisis de Valores Atípicos (Outliers)

- Manejo de Valores Atípicos
  - Los valores atípicos en las variables numéricas fueron identificados mediante cálculos del rango intercuartil
  - Los valores fuera del rango aceptable fueron tratados mediante winsorización o eliminados si eran inconsistentes con la lógica de los datos

1. Boxplot - Horas de Estudio



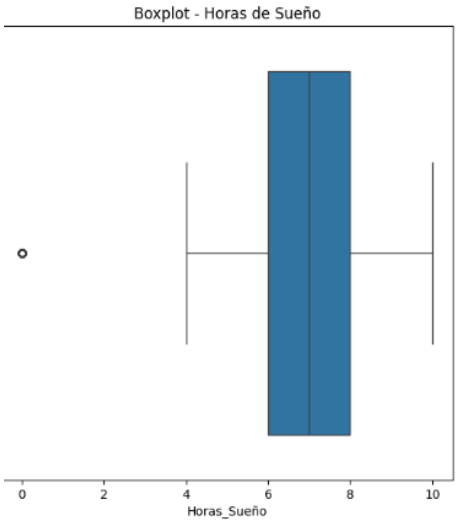
- Distribución:
  - El rango intercuartil (IQR) abarca la mayoría de los datos entre aproximadamente 15 y 25 horas.
    La mediana está alrededor de 20 horas, indicando que la mitad de los estudiantes estudian menos de ese tiempo.
- Valores Atípicos.
  - Existen algunos valores extremos por encima de las 40 horas, que podrían ser casos inusuales o errores de entrada.



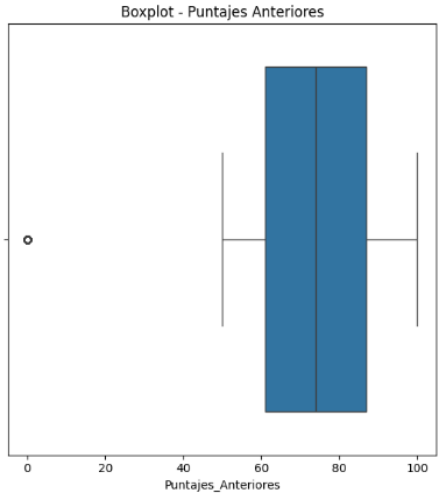
2. Boxplot - Asistencia

- Distribución:
  - El rango intercuartil (IQR) se encuentra entre el 60% y el 100% de asistencia.
  - La mediana está alrededor del 85%, lo que refleja que la mayoría de los estudiantes tienen una asistencia alta.
- Valores Atípicos:
  - Se identifican valores extremos menores al 20%, posiblemente correspondientes a estudiantes con asistencia irregular.

3. Boxplot - Horas de Sueño



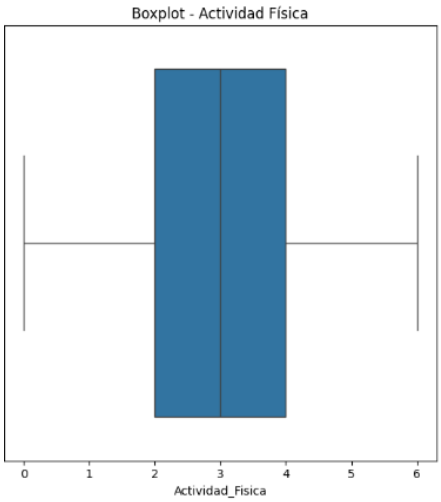
- Distribución:
  - La mayoría de los datos están dentro del rango intercuartil de 6 a 8 horas.
  - La mediana está cerca de 7 horas, lo que sugiere un patrón de sueño estable entre los estudiantes.
- Valores Atípicos:
  - Se observan valores extremos menores a 2 horas, que podrían reflejar casos severos de privación del sueño.



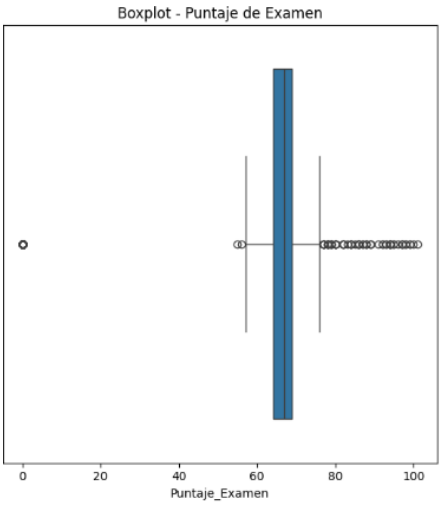
4. Boxplot - Puntajes Anteriores

- Distribución:
  - Los datos están concentrados entre 60 y 80 puntos, con un rango intercuartil claramente definido.
  - La mediana está cerca de 75 puntos, indicando un buen desempeño previo promedio.
- Valores Atípicos:
  - Existen valores extremos menores a 20 puntos, que podrían representar estudiantes con bajo rendimiento previo.

5. Boxplot - Actividad Física



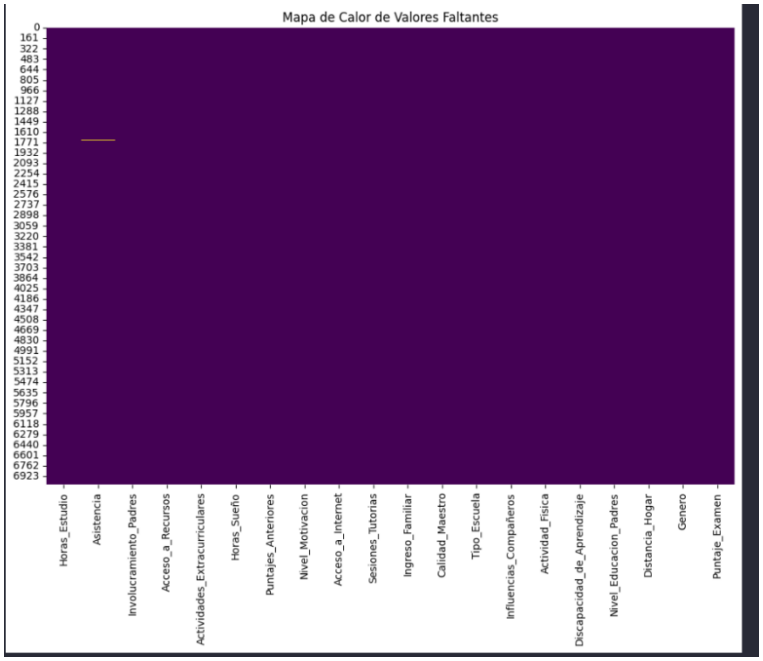
- Distribución:
  - El rango intercuartil abarca de 2 a 4 horas por semana.
  - La mediana está cerca de 3 horas, lo que sugiere un nivel moderado de actividad física.
- Valores Atípicos:
  - No se identifican valores extremos significativos en esta variable.



6. Boxplot - Puntaje de Examen

- Distribución:
  - Los datos se concentran entre 60 y 80 puntos, con una mediana cercana a 70 puntos, indicando un desempeño promedio positivo.
- Valores Atípicos:
  - Se observan múltiples valores extremos menores a 40 puntos, representando bajo desempeño, y algunos superiores a 95 puntos, posiblemente estudiantes con alto rendimiento.

5. Análisis de Valores Faltantes



Cada fila representa un registro en el dataset, y cada columna corresponde a una variable o característica específica. Los puntos amarillos indican valores faltantes, mientras que el color púrpura representa datos presentes.

En el eje horizontal se encuentran las variables del dataset, Estas variables parecen estar relacionadas con un análisis educativo, social o demográfico. En el eje vertical se representan los registros del dataset.

Los valores faltantes están localizados alrededor de las filas 1700 a 1900 y afectan una o más variables específicas, mientras que el resto del dataset está completamente lleno.

Esto indica que el dataset es mayormente completo, salvo por algunas excepciones concentradas en un subconjunto de registros.

Para analizar los valores faltantes, se deben identificar las columnas afectadas y evaluar el porcentaje de datos faltantes. Dependiendo de su impacto, se pueden aplicar técnicas como imputación, rellenado con medias, medianas o métodos avanzados, eliminación de registros, o simplemente ignorarlos si no afectan el análisis.

6. Observaciones y Hallazgos Importantes

La variable Puntajes_Anteriores tiene una correlación fuerte positiva con el Puntaje_Examen, lo que indica que los estudiantes con mejores resultados previos tienden a tener mejores puntajes en el examen actual, las Horas_Estudio también está positivamente correlacionada con el Puntaje_Examen, pero en menor medida, esto sugiere que estudiar más tiene un impacto moderado en el desempeño.

- Patrones interesantes:

Asistencia muestra una correlación moderada positiva con el puntaje, indicando que los estudiantes con mayor asistencia suelen tener mejores resultados, Sorprendentemente, Horas_Sueño tiene una correlación baja o incluso negativa, lo que puede indicar que otros factores influyen más en el rendimiento que las horas de sueño.

- Anomalías:

Algunas variables como Actividad_Física pueden mostrar correlación casi nula con el puntaje, lo que sugiere que no tienen un impacto directo en el desempeño académico y se pueden identificar outliers en Horas_Estudio y Puntaje_Examen, que podrían ser estudiantes con hábitos extremos o con resultados atípicos.

El análisis de correlación y los gráficos muestran qué variables tienen mayor impacto en la variable de interés. Esto guía decisiones para modelos predictivos o análisis posteriores.

En este caso:

Puntajes_Anteriores y Horas_Estudio son las variables más influyentes, mientras que otras como Actividad_Física tienen menor relevancia.

Las observaciones inusuales (outliers) podrían ser filtradas o investigadas más a fondo para entender su impacto.

Modelo de Machine Learning

- Modelo seleccionado: Regresión logística.
- Justificación: Este modelo es adecuado para predecir probabilidades, como la aprobación de un examen.

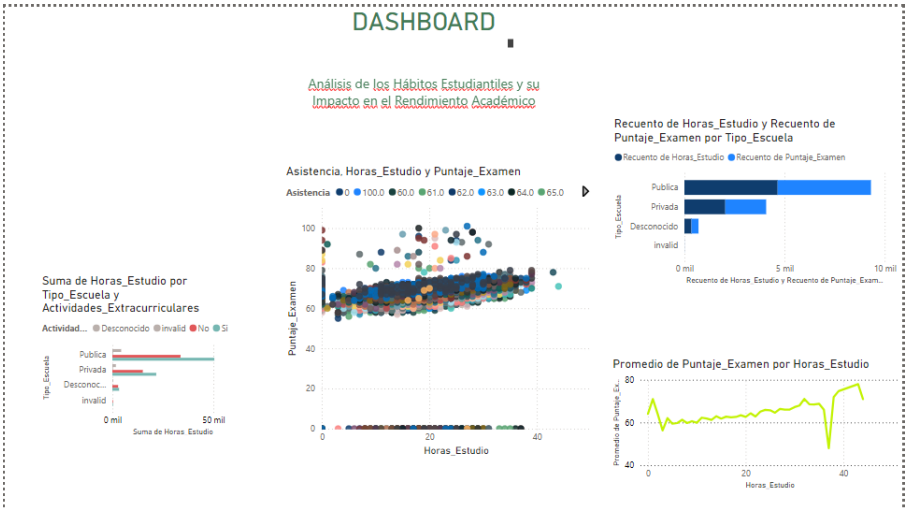
```
# Escalar las caracteristicas
scaler = StandardScaler()
X_train_scaled = scaler.fit_transform(X_train_smote)
X_test_scaled = scaler.transform(X_test)

# Entrenar el modelo de Regresión Logistica
modelo_lr = LogisticRegression(max_iter=1000, class_weight='balanced')
modelo_lr.fit(X_train_scaled, y_train_smote)

# Hacer predicciones
y_pred_lr = modelo_lr.predict(X_test_scaled)

# Evaluar el modelo
print("Resultados con Regresión Logistica:")
print(f'Precisión: {accuracy_score(y_test, y_pred_lr)}')
print("Reporte de clasificación:")
print(classification_report(y_test, y_pred_lr))
print("Matriz de Confusión:")
print(confusion_matrix(y_test, y_pred_lr))
```

```
Resultados con Regresión Logistica:
Precisión: 0.6024096385542169
Reporte de clasificación:
              precision    recall  f1-score   support

           0       0.06      0.46      0.11        79
           1       0.95      0.61      0.74      1332

    accuracy                           0.60      1411
   macro avg       0.51      0.53      0.43      1411
weighted avg       0.90      0.60      0.71      1411

Matriz de Confusión:
[[ 36  43]
 [518 814]]
```

- Implementación:
    - División de datos en conjuntos de entrenamiento (80%) y prueba (20%).

- Evaluación con métricas como precisión y recall.

DashBoard



1. Asistencia, Horas de Estudio y Puntaje de Examen:
    - Gráfico de dispersión: Muestra la relación entre las horas de estudio y el puntaje obtenido en el examen. Cada punto representa a un estudiante.Se puede observar que no existe una relación lineal clara entre ambas variables. Es decir, no se puede afirmar que a mayor cantidad de horas de estudio, necesariamente se obtenga un mejor puntaje.
2. Recuento de Horas de Estudio y Recuento de Puntaje de Examen por Tipo de Escuela:
    - Gráfico de barras: Compara la cantidad de horas de estudio y el puntaje obtenido en diferentes tipos de escuelas Permite identificar si existen diferencias significativas en los hábitos de estudio y el rendimiento académico entre los distintos tipos de escuelas.
3. Suma de Horas de Estudio por Tipo de Escuela y Actividades Extracurriculares:
    - Gráfico de barras: Muestra la cantidad total de horas de estudio por tipo de escuela y si el estudiante realiza o no actividades extracurriculares. Permite analizar si la participación en actividades extracurriculares influye en la cantidad de tiempo dedicado al estudio.
4. Promedio de Puntaje de Examen por Horas de Estudio:
    - Gráfico de línea: Muestra cómo varía el puntaje promedio del examen a medida que aumenta la cantidad de horas de estudio. Permite identificar si existe alguna tendencia general en el rendimiento académico en función del tiempo dedicado al estudio.

A partir de la información presentada en el dashboard, podemos inferir que:

- No existe una relación directa entre las horas de estudio y el rendimiento académico: Otros factores, como la calidad del estudio, las habilidades individuales y el tipo de escuela, también influyen en los resultados.
- Existen diferencias en los hábitos de estudio y el rendimiento académico entre los diferentes tipos de escuelas: Es necesario realizar un análisis más detallado para determinar las causas de estas diferencias.

- La participación en actividades extracurriculares puede tener un impacto en la cantidad de tiempo dedicado al estudio: Sin embargo, se requiere más información para determinar si esto afecta positivamente o negativamente al rendimiento académico.

Conclusiones y Futuras Líneas de Trabajo

- Hallazgos principales:
    - Hábitos como horas de estudio y motivación tienen mayor influencia en el éxito académico.
    - Factores externos como el apoyo familiar y la calidad del maestro también juegan un papel significativo.
- Mejoras futuras:
    - Ampliar la base de datos con información de más contextos escolares.
    - Explorar el uso de modelos más complejos, como árboles de decisión o redes neuronales.

Github

JaredCrz1806/ProyectoFinal