UAX Universidad Alfonso X el Sabio

ANÁLISIS DE FACTORES INFLUYENTES EN LA PERFORMANCE ACADÉMICA DE ESTUDIANTES

Camila Spinelli

Fundamentos de la Ciencia de Datos

1. Título del Proyecto

- "Análisis de factores influyentes en la performance de estudiantes aplicando metodología CRISP-DM".

2. Introducción – El Problema de Negocio

- Se desea desarrollar un modelo que ayude a identificar los patrones en el entorno y comportamiento de los estudiantes, de forma que se permita a la institución educativa y sus interesados a implementar estrategias para ayudar a mejorar el rendimiento académico.
- Algunas preguntas que se desean responder con el proyecto de ciencia de datos:
    - ¿Cuáles son las variables que tienen mayor nivel de influencia en el rendimiento académico?
    - ¿Cómo impacta el entorno familiar en los resultados de los estudiantes?
    - ¿Existen patrones en los hábitos de estudio que se relacionen con la performance del estudiante?
    - ¿Existe una relación entre los factores personales y la performance académica?

3. Objetivos

- Generar un modelo de datos que permita identificar y analizar que factores influyen en el rendimiento académico de los estudiantes, con el objetivo de obtener patrones que ayuden a poder proporcionar diferentes accionables para optimizar su performance en los exámenes.
- Se desea:
    - Evaluar el impacto de las variables de habito de estudio (pe: asistencia a clase, horas de estudio, sesiones de tutoría) del estudiantes en el rendimiento académico.
    - Analizar la influencia del contexto familiar en los resultados.
    - Monitorizar y visualizar mediante un dashboard el impacto de las diferentes variables clave en los resultados académicos.
    - Entregar un dashboard de PowerBI con un reporte de las conclusiones finales.

4. Fuentes de datos

- El DataSet que se va a utilizar en este proyecto contiene información de **6607 estudiantes** con **20 variables** que representan sus hábitos de estudio, entorno familiar, datos personales y resultados académicos.
    - Datos personales: genero, problemas de aprendizaje, horas de sueño, nivel de motivación, participación en actividades extracurriculares, influencia de pares.
    - Habito de estudio: Cantidad de horas dedicadas al estudio, asistencia a clase, asistencia a tutorías.
    - Entorno familiar: nivel de involucramiento familiar, nivel de educación de los padres, nivel económico familiar, acceso a internet.
    - Entorno académico: Acceso a recursos, calidad de profesor, tipo de educación.
- Las columnas incluyen tanto datos cuantitativos como cualitativos y los datos son de tipo numérico o string.
- En la siguiente pagina se adjunta dos tablas:
    - La primera con una descripción de los datos: nombre de columnas, tipo de datos, clasificación de medida y una breve descripción del dato.
    - La segunda con una descripción estadística de los datos cuantitativos.
- Estos datos son de tipo cross sectional ya que representan las características de los estudiantes en un único punto en el tiempo, de origen estructurado y de fuente primaria.
- Se puede encontrar el jupyter notebook y markdown del análisis inicial de datos en el repositorio de GitHub bajo la carpeta de informes.

Descripción general de los datos:

Columna	Tipo de dato	Clasificación	Breve descripción
Hours_studied	Int64	Cuantitativo	Horas dedicadas al estudio
Attendance	Int64	Cuantitativo	Porcentaje de asistencia a clase
Parental_involment	String	Cualitativo	Nivel de involucramiento de los padres (bajo, medio, alto)
Access_to_Resources	String	Cualitativo	Nivel de acceso a recursos (bajo, medio, alto)
Extracurricular_Activities	String	Cualitativo	Estudiante realiza actividades extracurriculares (si, no)
Sleep_Hours	Int64	Cuantitativo	Horas de sueño
Previous_Scores	Int64	Cuantitativo	Resultados anteriores
Motivation_Level	String	Cualitativo	Nivel de motivación de estudio (bajo, medio, alto)
Internet_Access	String	Cualitativo	Estudiante cuenta con acceso a internet (si, no)
Tutoring_Sessions	Int64	Cuantitativo	Sesiones de tutoria que el estudiante asistió
Family_Income	String	Cualitativo	Ingreso económico del núcleo familiar (bajo, medio, alto)
Teacher_Quality	String	Cualitativo	Calidad de profesor del estudiante (bajo, medio, alto)
School_Type	String	Cualitativo	Tipo de institución al que asiste (pública, privada)
Peer_Influence	String	Cualitativo	Nivel de influencia de sus pares (negativa, neutra, positiva)
Physical_Activity	Int64	Cuantitativo	Horas de actividad fisica
Learning_Disabilities	String	Cualitativo	Dificultad de aprendizaje (si, no)
Parental_Education_Level	String	Cualitativo	Nivel de educación de los padres (secundario, universitario, posgrado)
Distance_from_Home	String	Cualitativo	Distancia de la casa al centro educativo (lejos, cerca, moderado)
Gender	String	Cualitativo	Genero del estudiantes (femenino, masculino)
Exam_Score	Int64	Cuantitativo	Resultado final del exámen

Descripción estadística de los datos cuantitativos:

Columna	Media	Desviación Estándar	Min.	25 %	Mediana (50%)	75 %	Máx.
Hours_studied	19.98	5.99	1	16	20	24	44
Attendance	79.98	11.55	60	70	80	90	100
Sleep_Hours	7.03	1.47	4	6	7	8	10
Previous_Scores	75.07	14.40	50	63	75	88	100
Tutoring_Sessions	1.49	1.23	0	1	1	2	8
Physical_Activity	2.97	1.03	0	2	3	4	6
Exam_Score	67.24	3.89	55	65	67	69	101

5. Plan de Trabajo

Fase del Proyecto	Semana									
	01	02	03	04	05	06	07	08	09	10
Fase 1: Comprensión del negocio										
Fase 2: Comprensión de los datos										
Fase 3: Preparación de los datos										
Fase 4: Modelado										
Fase 5: Evaluación										
Fase 5: Implantación										
Entrega del Proyecto										

6. Metodología

- La metodología que se utilizara en este proyecto será CRIP-DM.
- Para responder a las preguntas plantadas se utilizaran modelos de regresión por la naturaleza de los datos a analizar.
- Para el entregable del proyecto se va a realizar un reporte final y un dashboard en PowerBI con gráficos y métricas clave para entender la influencia de cada variable.

7. Tecnología

- Gestión de datos: MySQL.

- Lenguaje de programación librerías: Python con Pandas.
- Plataforma de desarrollo del proyecto: Jupyter Notebooks.
- Control de versiones: GitHub.

8. Lectura y entendimiento de datos

- El script con la creación de la BBDD, la carga y las queries se encuentra en el repositorio de GitHub.
- Se adjunta captura de pantalla del script y de algunas de las salidas de respuesta de las queries.

```sql
DROP TABLE students;

CREATE TABLE students (
id INT PRIMARY KEY AUTO_INCREMENT,
hours_studied INT,
attendance INT,
parental_involvement VARCHAR(10),
access_to_resources VARCHAR(10),
extracurricular_activities VARCHAR(10),
sleep_hours INT,
previous_scores INT,
motivation_level VARCHAR(10),
internet_access VARCHAR(10),
tutoring_sessions INT,
family_income VARCHAR(10),
teacher_quality VARCHAR(10),
school_type VARCHAR(24),
peer_influence VARCHAR(24),
physical_activity INT,
learning_disabilities VARCHAR(10),
parental_education_level VARCHAR(24),
distance_from_home VARCHAR(24),
gender VARCHAR(10),
exam_score INT
);

--la carga de datos se hizo utilizando la opcion de importar datos de dbeaver
--tambien se podria usar el comando LOAD DATA INFILE 'path/to/csv' INTO table students

SELECT * FROM students s;
--Q1
SELECT id, attendance ,sleep_hours , hours_studied, previous_scores,
exam_score FROM students s WHERE exam_score > 85;
--Q2
SELECT AVG(exam_score) FROM students s WHERE attendance > 95;
--Q3
SELECT AVG(exam_score) FROM students s WHERE internet_access = 'No';
--Q4
SELECT COUNT(*) FROM students s WHERE (exam_score < previous_scores);
--Q5
SELECT id, attendance, hours_studied, tutoring_sessions, sleep_hours, previous_scores,
exam_score FROM students s WHERE (exam_score < previous_scores);
--Q6
SELECT tutoring_sessions , AVG(exam_score) from students s GROUP BY tutoring_sessions;

```

Query1:

```
"id","attendance","sleep_hours","hours_studied","previous_scores","exam_score"
95,89,4,18,73,100
218,70,7,19,54,89
405,77,5,17,53,86
530,83,7,15,97,97
771,96,6,24,93,94
837,76,8,29,96,94
920,74,6,21,94,97
1108,77,6,14,75,89
1110,69,7,31,52,92
1526,98,6,27,93,101
1608,98,9,30,93,88
1845,92,9,21,58,89
2293,70,9,21,66,91
2422,90,9,27,52,86
2426,83,4,23,89,99
2514,86,8,18,60,88
2596,69,8,7,54,87
2688,71,8,11,55,87
2905,62,7,11,76,88
3125,90,5,19,90,94
3142,63,10,7,90,86
3365,76,6,16,63,86
3458,93,7,18,76,96
3580,90,8,14,86,99
4193,90,9,28,91,98
4298,67,6,21,88,95
4406,98,7,25,90,94
4532,69,7,26,95,93
4584,73,7,25,56,93
4780,88,4,1,72,92
5967,99,7,25,77,97
6348,96,4,28,98,98
6394,83,8,16,92,98
6523,90,6,18,54,95
```

Query2:

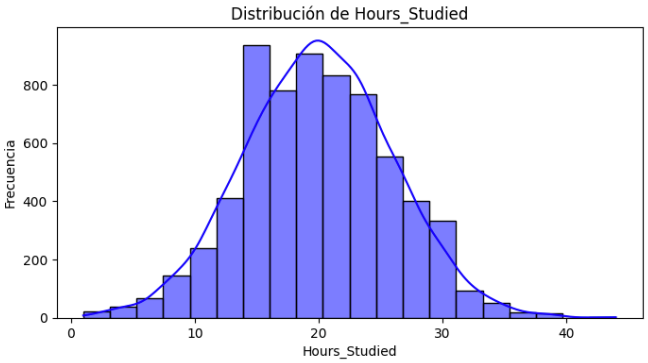
```
"AVG(exam_score)"
70.6445

```

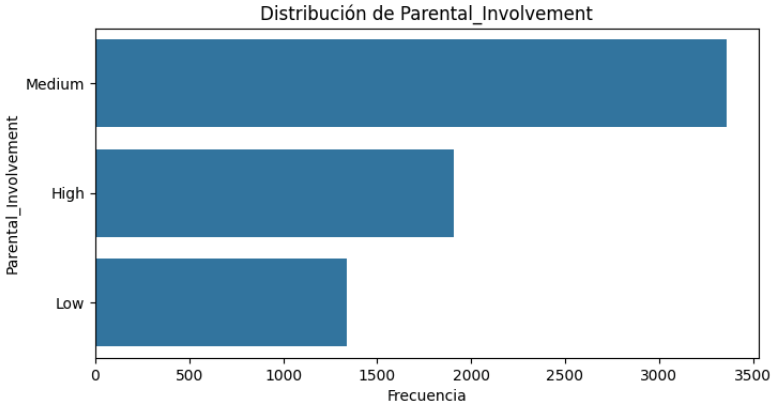
Query6:

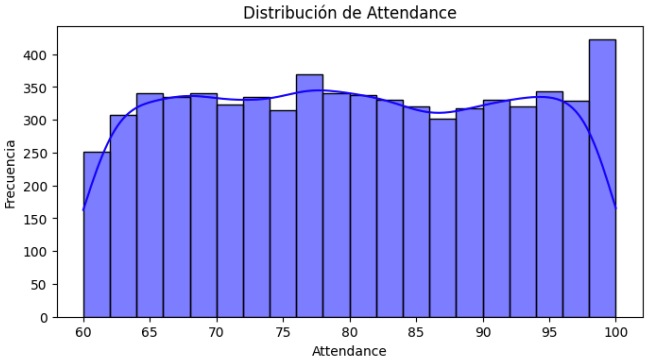
```
"tutoring_sessions","AVG(exam_score)"
0,66.4898
2,67.5670
1,66.9803
3,67.8947
4,68.2292
5,69.0680
6,71.6667
7,69.8571
8,69.0000

```

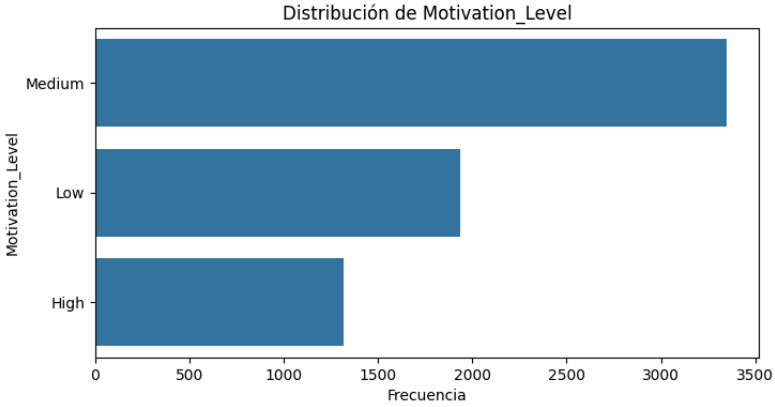
9. Análisis Exploratorio de Datos

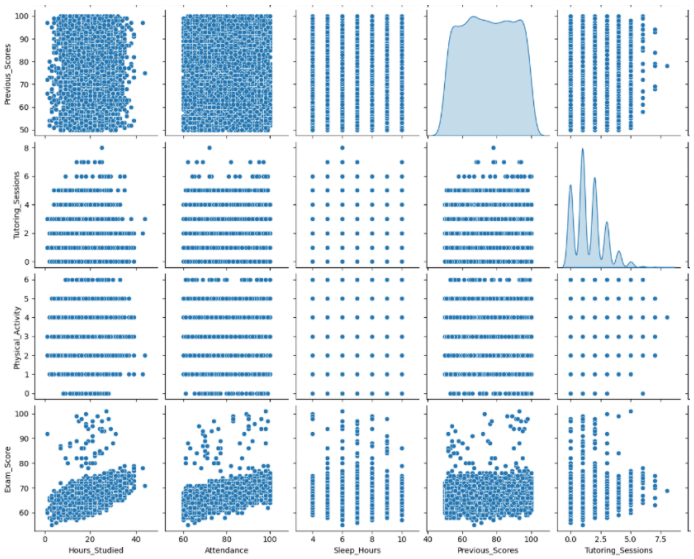
- Se adjunta en GitHub el HTML del notebook con los resultados de los análisis univariados y bivariados y a continuación algunas capturas de pantalla representativas de las gráficas obtenidas en los análisis:

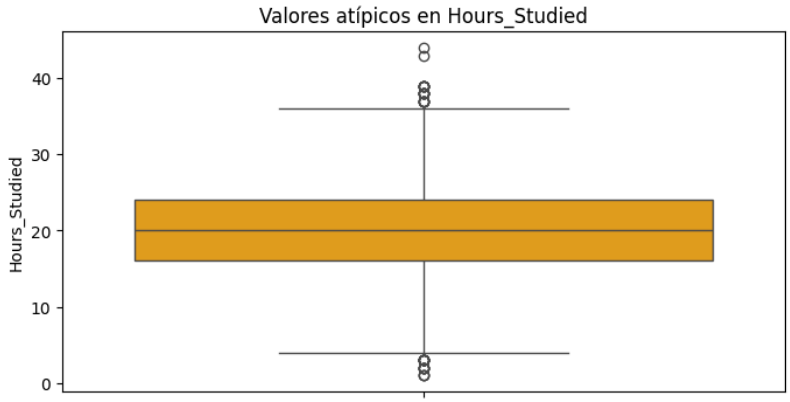






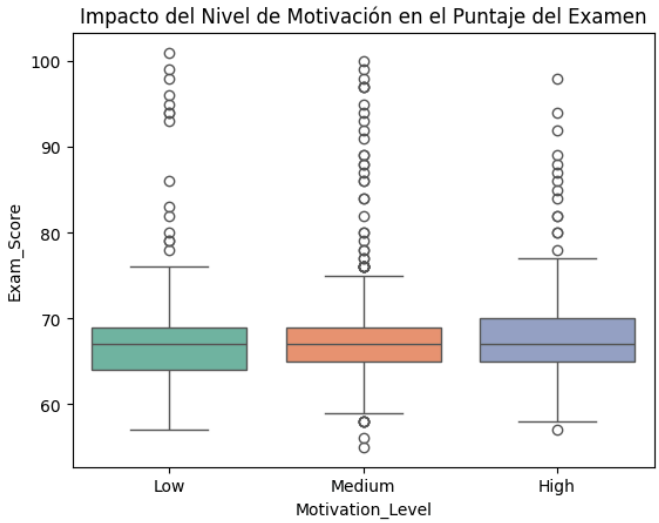




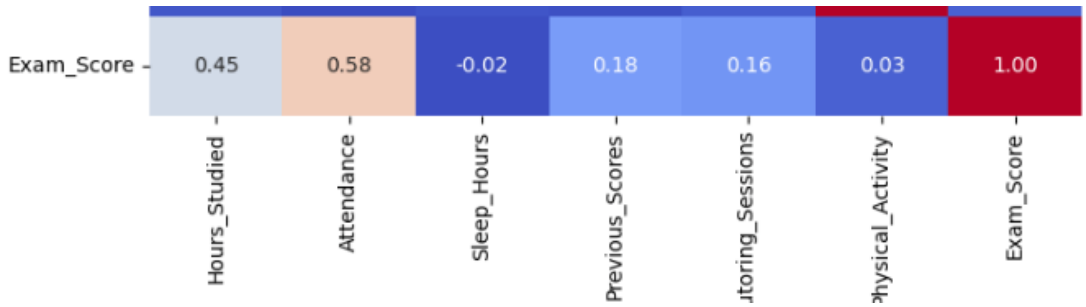


Patrones observados:

- Distribuciones:
  - Horas de Estudio (Hours_Studied): La mayoría de los estudiantes entre 16 y 24 horas por semana. Lo que coincide con valor 19 que obtuvimos de calcular la media. Existen valores extremos tanto superiores como inferiores (Min. de 1 hora y Max. De 44 horas)
  - Horas de sueño (Sleep_Hours): La distribución es bastante centrada. La mayoría de los estudiantes duermen entre 6 y 8 horas por noche. Se detectan algunos extremos (menos de 4 horas, y mas de 10 horas) que podrían influir de forma negativa en los resultados.
- Relaciones identificadas:
  - Impacto del nivel de Motivación:
    - Los estudiantes con niveles de motivación altos tiene mejores resultados académicos en los exámenes. Mientras que los que tiene motivación baja tienen una mayor variabilidad, con algunos estudiantes mostrando puntajes muy altos.



  - Se observa que, aunque es muy débil, parece haber una relación positiva entre la actividad física y los resultados de exámenes. Los estudiantes con niveles mas altos de actividad física tienden a obtener puntajes ligeramente mejores. Esto podría darse a un estilo de vida equilibrado entre estudio y actividades extracurriculares o los beneficios del ejercicio como la reducción del estrés.
  - Se confirma que las sesiones de tutorías, la asistencia a clase y la cantidad de horas de estudio son beneficiosas para mejores resultados.
  - Por otro lado, los estudiantes que reportaron una calidad de enseñanza alta obtienen puntajes consistentemente mas altos. Lo cual destaca la importancia de los educadores en la performance.



Problemas detectados:

- Outliers:
  - Exam_Score presenta valores por encima de 100, lo cual no es lógico en una escala de 0 a 100.
  - Para Hours_Studied algunos estudiantes tiene valores muy altos o muy bajos (44 horas por semana).
- Valores Faltantes:
  - Teacher_Quality presenta 78 datos nulos, Parental_Education_Level tiene 90 y Distance_from_Home cuenta con 67 datos nulos.
  - Estos datos nulos representan un porcentaje muy bajo del total de los datos.

9. ¿Qué nos dicen los datos?

- La mayoría de los estudiantes estudian entre 16 y 24 horas por semana y duermen entre 6 y 8 horas por noche. Estos hábitos parecen estar relacionados con puntajes promedio en los exámenes.
- Los estudiantes con alto nivel de motivación tienden a obtener mejores puntajes, aunque algunos con baja motivación logran destacar. Esto nos indica que hay otros factores que influyen en la performance.
- Existe una relación moderada entre las horas de estudio y los resultados finales. Los estudiantes que estudian mas horas suelen obtener mejores puntajes.
- La calidad del profesorado muestra un impacto significativo en el rendimiento. Los estudiantes que perciben a sus profesores como de alta calidad tienen un mejor rendimiento.
- La actividad física tiene una relación débil pero positiva en el rendimiento académico, sugiriendo que un estilo de vida equilibrado favorece al aprendizaje.
- Por lo tanto, las variables Motivation_Level, Teacher_Quality y Hours_Studied son factores claves que afectan el rendimiento académico. Un enfoque en mejorar estos aspectos podría ayudar con la performance de los estudiantes.

UAX Universidad Alfonso X el Sabio

G R A C I A S

UAX.COM