



# UNIVERSIDAD DE COLIMA

## **Diplomado: Analítica y Ciencia de datos** **Proyecto Final**

### **Integrantes**

- Akary Larios Pulido
- Luis Eduardo Oropeza Garcia
- Luis Jesus Cardenas Peregrino
- Fernando Martin Rodriguez Moreno
- María Valeria García Verdugo
- Emiliano Cortés Ortega

### **1.- Selección y Justificación de la Base de Datos**

#### **Base de datos elegida:**

“ student performance factors” / factores de rendimiento estudiantil

<https://www.kaggle.com/datasets/lainguyn123/student-performance-factors>

Este conjunto de datos nos ofrece una comprensiva revisión de varios factores sociales y educativos que influyen en los resultados de los alumnos durante la presentación de exámenes; a la vez nos habla de hábitos, asistencia, el nivel de compromiso de los padres y otros aspectos que pueden ayudar a determinar razones para un éxito académico dentro de un plantel o sistema educativo.

#### **Justificación**

Dentro de la discusión y búsqueda por un tema de interés común para el equipo, decidimos optar por algo relacionado con la educación. Los sistemas educativos tanto públicos como privados a menudo se modifican con el propósito de una

optimización de resultados en las pruebas de aprovechamiento escolar; esto para poder solicitar mejoras en escuelas así como recursos; y aunque este por sí mismo es un derecho, es cierto que también existe una contienda interminable por mejorar los promedios del alumnado para representar la calidad de las instituciones, por lo que sí es de suma importancia saber cómo y porqué se obtienen buenos, malos o regulares resultados.

El equipo concuerda con que no solo es de interés para las escuelas en particular, sino también para el diseño más asertivo de políticas públicas enfocadas al sector, para resolver inconformidades sociales, y mejorar el ambiente para un fundamental futuro pilar de nuestra sociedad; los jóvenes.

Mediante las facilidades gratuitas que se ofrecen en “Kaggle”, se logró encontrar una base de datos limpia y con elementos que ayudan a nuestro análisis a modo de práctica y de utilidad para el proyecto. La data disponible es un recurso abierto y de simulación que nos ofrece un preámbulo adecuado para obtener los objetivos deseados del proyecto.

### **¿Qué preguntas quiero responder con este análisis?**

¿Qué factores influyen en las mejores calificaciones de los alumnos?

¿Qué factores influyen en una baja calificación?

¿Qué nivel de aprovechamiento es más común?

¿Cómo podemos mejorar el panorama actual de resultados en estudiantes con calificaciones más bajas?

### **Factores de rendimiento de los estudiantes**

El rendimiento académico es un gran tema de interés y preocupación a nivel mundial ya que en esta etapa llega tener un impacto significativo en el desarrollo personal de cada individuo de una sociedad.

La educación es un derecho fundamental de todas las personas y es la clave para la construcción de sociedades más justas y equitativas, ya que se considera el mejor factor de producción esta nos permite acabar de raíz con muchos de los problemas económicos de una nación y funge como instrumento regulador de las desigualdades sociales, ***“saber más para servir mejor”***.

Partiendo de esto al elegir este conjunto de datos, el cual nos proporciona una descripción general de varios factores que afectan el rendimiento de los estudiantes en los exámenes, incluyendo información sobre los hábitos de estudio, la asistencia, la participación de los padres y otros aspectos que influyen en el éxito académico, buscamos dar a conocer los principales factores que necesitamos mitigar tanto en el personal docente como familiar, también cuantificar las necesidades para el solicitar recursos gubernamentales.

## 2.- Preparación y Limpieza de los Datos

### Descripción de la Base de Datos:

La base de datos cuenta con 6607 registros y 20 campos los cuales se describen a continuación:

Atributo	Descripción	Tipo de Dato
Hours_Studied	Número de horas dedicadas al estudio por semana.	Integer
Attendance	Porcentaje de clases asistidas.	Float
Parental_Involvement	Nivel de implicación de los padres en la educación del alumno (Low, Medium, High).	String
Access_to_Resources	Disponibilidad de recursos educativos (Low, Medium, High).	String
Extracurricular_Activities	Participación en actividades extracurriculares (Yes, No).	String
Sleep_Hours	Número de horas promedio de sueño por noche.	Integer
Previous_Scores	Puntuaciones de exámenes anteriores.	Integer
Motivation_Level	Nivel de motivación del alumno (Low, Medium, High).	String
Internet_Access	Disponibilidad de acceso a internet (Yes, No).	String
Tutoring_Sessions	Número de sesiones de tutoría a las que se asiste al mes.	Integer
Family_Income	Nivel de ingresos familiares (Low, Medium, High).	String
Teacher_Quality	Calidad de los profesores (Low, Medium, High).	String
School_Type	Tipo de escuela a la que se asiste (Public, Private).	String
Peer_Influence	Influencia de los compañeros en el rendimiento académico (Positive, Neutral, Negative).	String
Physical_Activity	Número de horas promedio de actividad física a la semana.	Integer
Learning_Disabilities	Presencia de problemas de aprendizaje (Yes, No).	String
Parental_Education_Level	Nivel educativo más alto de los padres (High School, College, Postgraduate).	String
Distance_from_Home	Distancia de casa a la escuela (Near, Moderate, Far).	String
Gender	Género del estudiante (Male, Female).	String
Exam_Score	Nota del examen final.	Integer

### Limpieza de Datos:

- Identificación y manejo de valores faltantes
- Corrección de errores en los datos
- Transformaciones necesarias

Descripción:

Haciendo uso de la función de Excel CONTAR.BLANCO, nos percatamos que el 3.47% de los registros contaban con al menos un campo en blanco (null), ya que no se nos hizo relevante decidimos eliminar los registros.

Al final de cada fila colocamos la función antes mencionada, con el rango de todos los campos, corrimos la fórmula para duplicar en cada uno de los registros, con el uso de un filtro seleccionamos los registros mayores a cero y procedimos a eliminar las filas.

### 3. Análisis Exploratorio de Datos (EDA)

#### Análisis Descriptivo:

- Estadísticas descriptivas (media, mediana, moda, desviación estándar)
- Distribuciones de variables/campos.

HORAS DE ESTUDIO	CALIFICACION PROMEDIO	ESTUDIANTES	%
1-10	64	365	6%
11-20	66	3,179	48%
21-30	69	2,810	43%
31-40	71	251	4%
41-50	75	2	0%

Hours_Studied	
PROMEDIO	20
MODA	20
DESVIACION	6

% CLASES ASISTIDAS	CALIFICACION PROMEDIO	ESTUDIANTES	%
60-69	64	1,573	24%
70-79	66	1,681	25%
80-89	68	1,607	24%
90-100	70	1,746	26%

Attendance	
PROMEDIO	80
MODA	67
DESVIACION	12

IMPLICACIÓN DE LOS PADRES	CALIFICACION PROMEDIO	ESTUDIANTES	%
Low	66	1,337	20%
Medium	67	3,362	51%
High	68	1,908	29%

DISPONIBILIDAD DE RECURSOS	CALIFICACION PROMEDIO	ESTUDIANTES	%
Low	66	1,313	20%
Medium	67	3,319	50%
High	68	1,975	30%

Extracurricular Activities	CALIFICACION PROMEDIO	ESTUDIANTES	%
No	67	2,571	40%
Yes	67	3,807	60%

Motivacion	CALIFICACION PROMEDIO	ESTUDIANTES	%
High	68	1,277	20%
Low	67	1,864	29%
Medium	67	3,237	51%

Acceso a internet	CALIFICACION PROMEDIO	ESTUDIANTES	%
No	67	485	
Yes	67	5,893	

Horas de Sueño	CALIFICACION PROMEDIO	ESTUDIANTES	%
4	68	296	5%
5	67	668	10%
6	67	1,322	21%
7	67	1,682	26%
8	67	1,354	21%
9	67	753	12%
10	67	303	5%

Horas de Sueño	
PROMEDIO	7
MODA	7
DESVIACION	1

Ingresos Familiares	CALIFICACION PROMEDIO	ESTUDIANTES	%
Low	67	2,582	40%
Medium	67	2,566	40%
High	68	1,230	19%

Teacher Quality	CALIFICACION PROMEDIO	ESTUDIANTES	%
Low	67	647	10%
Medium	67	3,826	60%
High	68	1,905	30%

School Type	CALIFICACION PROMEDIO	ESTUDIANTES	%
Private	67	1,944	30%
Public	67	4,434	70%

Tutoring Sessions	CALIFICACION PROMEDIO	ESTUDIANTES	%
0	67	1,458	23%
1	67	2,111	33%
2	68	1,586	25%
3	68	800	13%
4	68	296	5%
5	69	101	2%
6	72	18	0%
7	70	7	0%
8	69	1	0%

Tutoring Sessions	
PROMEDIO	1.5
MODA	1.0
DESVIACION	1.2

Peer Influence	CALIFICACION PROMEDIO	ESTUDIANTES	%
Negative	67	1,330	21%
Neutral	67	2,495	39%
Positive	68	2,553	40%

Learning Disabilities	CALIFICACION PROMEDIO	ESTUDIANTES	%
No	67	5,710	90%
Yes	66	668	10%

Parental Education	CALIFICACION PROMEDIO	ESTUDIANTES	%
College	67	1,939	30%
High School	67	3,159	50%
Postgraduate	68	1,280	20%

Distancia de Casa	CALIFICACION PROMEDIO	ESTUDIANTES	%
Far	66	636	10%
Moderate	67	1,941	30%
Near	68	3,801	60%

Physical Activity	CALIFICACION PROMEDIO	ESTUDIANTES	%
0	68	44	1%
1	67	400	6%
2	67	1,562	24%
3	67	2,460	39%
4	67	1,535	24%
5	68	346	5%
6	68	31	0%

Horas actividad Física	
PROMEDIO	3.0
MODA	3.0
DESVIACION	1.0

Genero	CALIFICACION PROMEDIO	ESTUDIANTES	%
Female	67	2,690	42%
Male	67	3,688	58%

Datos para sacar la media, mediana, moda : resultados de exámenes, horas estudiadas, horas de asistencia, horas de sueño.

## 4. Visualización de Datos

• **Herramienta de Visualización:** Uso de Tableau para crear visualizaciones interactivas

• **Software de análisis de datos:** Excel

• **Principales Visualizaciones:**

[https://public.tableau.com/app/profile/luis.cardenas4701/viz/ProyectoFinal\\_Tableau\\_17283596205450/Historia](https://public.tableau.com/app/profile/luis.cardenas4701/viz/ProyectoFinal_Tableau_17283596205450/Historia)

## 5. Interpretación y Conclusiones

• **Resultados del Análisis:**

- El promedio de calificación por género no muestra gran diferencia 67.27 para mujeres y 67.23 para hombres
- El factor escuela privada o pública no es de gran relevancia para una calificación alta o baja
- El acceso a internet es un factor de gran impacto para obtener buena calificación
- El apoyo de los padres de medio a alto contribuye positivamente a una mejor calificación
- El acceso a recursos estudiantiles de medio a alto contribuye positivamente a una mejor calificación
- Los estudiantes que dedicaron más de 15 horas de estudio muestran calificaciones más altas

• **Conclusiones:**

• **Conclusiones generales del proyecto**

• El promedio general de examen para los estudiantes de la base de datos analizada fue de 67.35

• Los factores de mayor relevancia observados para este análisis fueron acceso a internet, acceso a recursos estudiantiles, apoyo de los padres y horas de estudio. La

presencia de los recursos mencionados en un nivel alto o medio favorece una mayor calificación y viceversa.

- No existe una fórmula mágica o verdad absoluta ya que también se observó una minoría estudiantes con limitaciones o carencias de estos recursos obteniendo buena calificación, claro que la probabilidad de que esto suceda es menor.

#### **• Recomendaciones basadas en los resultados obtenidos**

- Hacer lo posible para tener acceso a internet así como otros recursos estudiantiles
- Llevar una buena relación con la familia y apoyarse en ella
- Estudiar al menos 15 horas

#### **• Posibles limitaciones del análisis y sugerencias para trabajos futuros**

- La base de datos elegida muestra un porcentaje muy pequeño de estudiantes con calificaciones altas lo cual dificulta el análisis y recomendaciones ya que observamos estudiantes en igualdad de recursos y calificaciones menores
- A futuro se podría buscar una base de datos que incluya escuela de procedencia así como estado/país para obtener un mejor análisis