

Trabajo Práctico Big Data – Grupo 4

Análisis y Modelado de Datos con Orange Data Mining

Dataset: Students Performance

1. Definición del Problema

“Queremos predecir si un estudiante tendrá un rendimiento académico alto o bajo, en base a sus hábitos de estudio, asistencia y apoyo familiar. Esto permitiría diseñar políticas educativas preventivas para mejorar el desempeño.”

2. Análisis Exploratorio de Datos

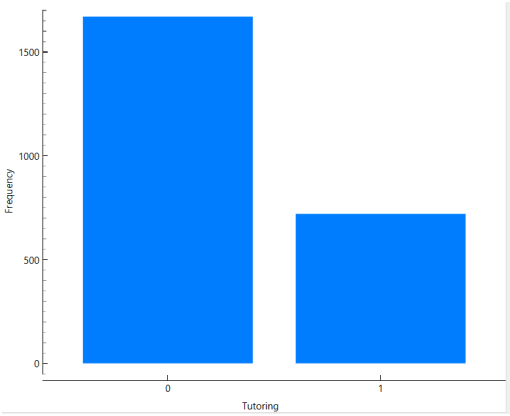
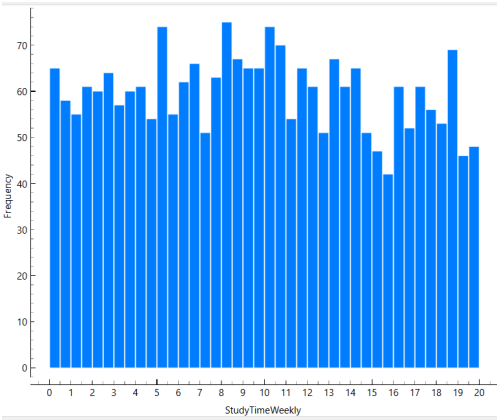
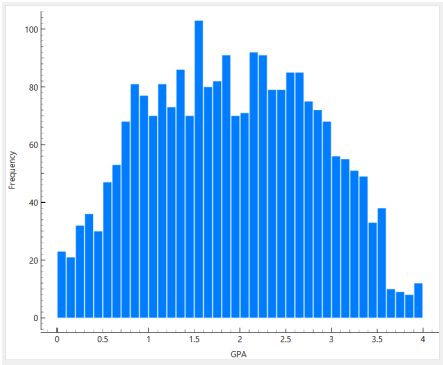
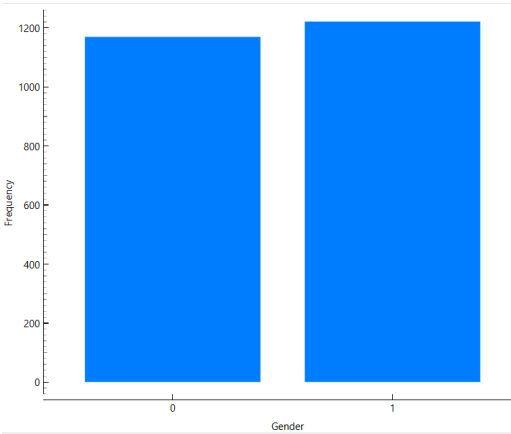
Descripción general del dataset

- Total de observaciones: 2392 estudiantes.
- Variables numéricas y categóricas relacionadas con edad, asistencia, hábitos de estudio y apoyo familiar.
- Variable objetivo: GPA (promedio académico).
- Utilizando Feature analytics y Distributions, obtuvimos un pantallazo inicial sobre media, mediana, moda, mínimos y máximos.

Feature statistics

	Name	Distribution	Mean	Mode	Median	Dispersion	Min.	Max.
N	StudentID		2196.50	1001	2196.50	0.31	1001	3392
N	Age		16.47	15	16	0.07	15	18
N	Ethnicity		0.88	0	0	1.17	0	3
N	ParentalEducati...		1.75	2	2	0.57	0	4
N	StudyTimeWee...		9.77199	0.00105654	9.70536	0.578346	0.00105654	19.9781
N	Absences		14.54	13	15	0.58	0	29
N	ParentalSupport		2.12	2	2	0.53	0	4
N	GPA		1.90619	0.00	1.89339	0.479997	0.00	4

Distributions



Nuestras variables

	Name	Type
1	StudentID	N numeric
2	Age	N numeric
3	Gender	C categorical
4	Ethnicity	N numeric
5	ParentalEducati...	N numeric
6	StudyTimeWee...	N numeric
7	Absences	N numeric
8	Tutoring	C categorical
9	ParentalSupport	N numeric
10	Extracurricular	C categorical
11	Sports	C categorical
12	Music	C categorical
13	Volunteering	C categorical
14	GPA	N numeric
15	GradeClass	N numeric

Algunos datos sobre nuestras variables:

StudentID

- Variable numérica usada únicamente como identificador de cada estudiante.
- No tiene valor analítico para el modelado.

Age

- Edad de los estudiantes, entre 15 y 18 años.
- Rango etario acotado que representa un grupo adolescente homogéneo.

Gender

- Codificada en 0 y 1 (masculino/femenino).
- Distribución bastante equilibrada entre ambos géneros.

Ethnicity

- Codificada de 0 a 3 (cuatro grupos).

- Cada número representa un grupo étnico distinto (clasificación anónima definida por el creador del dataset).
- No hay jerarquía entre categorías: es simplemente una variable categórica nominal.

ParentalEducation

- Codificada de 0 a 4 (niveles crecientes de educación de los padres).
- La mayoría de los valores se concentra en niveles intermedios.
- Posible variable asociada al rendimiento académico.

StudyTimeWeekly

- Horas de estudio semanal (0.00 a 19.97 horas).
- Promedio de 9.8 horas semanales con alta dispersión.
- Es una variable numérica clave para analizar hábitos de estudio.

Absences

- Cantidad de inasistencias (0 a 29).
- Promedio de 14,5 faltas, con algunos valores extremos.
- Alta probabilidad de correlación negativa con el rendimiento académico.

Tutoring

- Codificada en 0 y 1.
- Alrededor del 30 % de los estudiantes recibe tutorías.
- Posible factor de apoyo académico adicional.

ParentalSupport

- Codificada de 0 a 4.
- Promedio de 2,1, concentrado en niveles medios.
- Puede tener impacto positivo en el desempeño de los estudiantes.

Extracurricular

- Codificada en 0 y 1.
- 38 % de los estudiantes participa en actividades extracurriculares.

Sports

- Codificada en 0 y 1.
- 30 % practica deportes.

Music

- Codificada en 0 y 1.
- 20 % participa en actividades musicales.

Volunteering

- Codificada en 0 y 1.
- 16 % realiza actividades de voluntariado.

GPA

- Promedio general de notas sobre 4.0.
- Media de 1,9 → indica un rendimiento académico medio-bajo.
- Es una variable **clave como objetivo de modelado (target)** para regresión o clasificación.

GradeClass

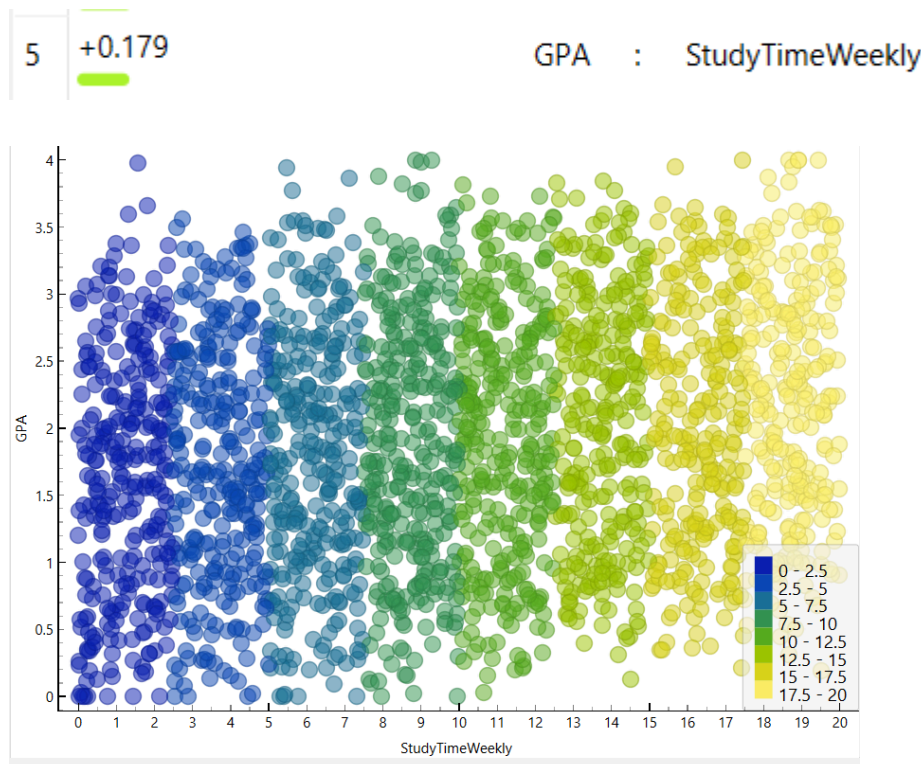
- Variable ordinal de 0 a 4 que agrupa niveles de rendimiento académico.
- Permite transformar el problema en clasificación si se desea.

Hipótesis sobre variables relevantes

Hipótesis 1 — Horas de estudio (StudyTimeWeekly)

“Los estudiantes que dedican más horas semanales al estudio tienden a obtener un **GPA más alto**.”

Razonamiento: mayor tiempo de estudio suele asociarse a mejores resultados académicos.



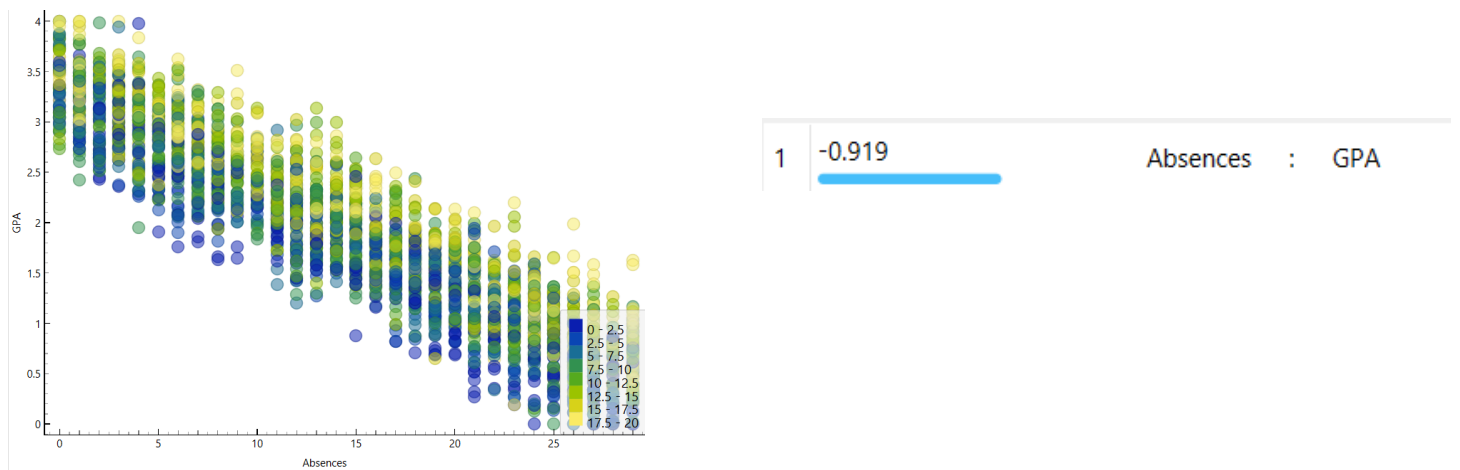
El gráfico de dispersión muestra una **tendencia ascendente**: a medida que aumenta el número de horas semanales de estudio, también se incrementan los valores de **GPA**.

La **correlación positiva** ($r = 0.179$) respalda esta tendencia, aunque no es extremadamente fuerte.

Hipótesis 2 — Inasistencias (Absences)

“Un mayor número de inasistencias se asocia a **GPA más bajo**.”

Razonamiento: faltar a clases reduce la exposición al contenido y el rendimiento.



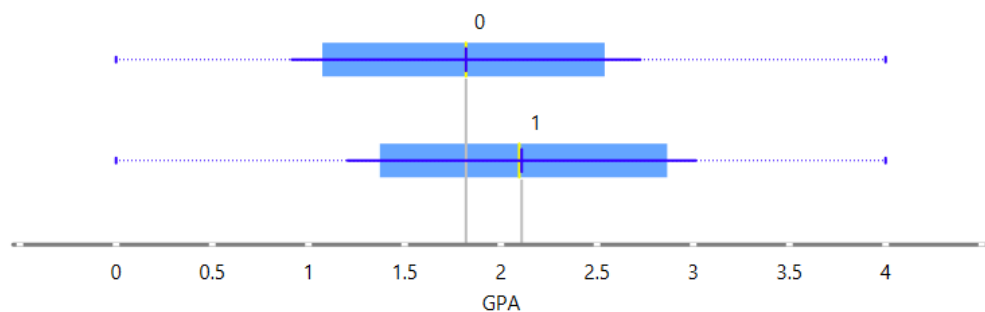
En este caso, el gráfico muestra una **tendencia descendente clara**: al aumentar la cantidad de inasistencias, los valores de **GPA** tienden a disminuir.

La **correlación negativa fuerte** ($r = -0.919$) respalda esta relación de manera contundente, indicando que la asistencia es un factor altamente asociado al rendimiento académico.

Hipótesis 3 — Apoyo parental (Tutoring**) —> No usamos la variable **ParentalSupport** ya que deberíamos transformarla a categorica y eso se hace en los próximos puntos, por eso usamos tutoring que de alguna manera expresa el apoyo de los padres.**

“Los estudiantes que reciben tutorías académicas (**Tutoring**) presentan un GPA promedio más alto que aquellos que no reciben tutorías.”

Razonamiento: el refuerzo académico suele mejorar el desempeño escolar.



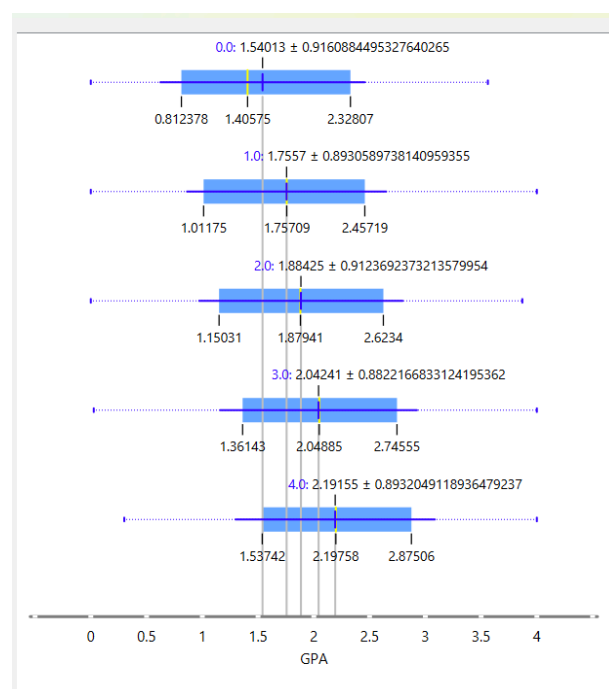
El gráfico muestra que los estudiantes que reciben tutorías académicas (valor 1) presentan un GPA promedio superior (alrededor de 2.0) en comparación con quienes no reciben tutorías (valor 0, promedio cercano a 1.5). Aunque la diferencia no es extrema, la tendencia es consistente y sugiere que el acompañamiento académico adicional tiene un impacto positivo en el rendimiento.

Hipótesis adicional — Apoyo parental (Parental Support**)**

No se utilizó esta variable en la etapa inicial porque requería ser transformada a categórica para poder analizarla. Sin embargo, era la que más se relacionaba con el problema planteado, ya que representaba el apoyo familiar dentro de los factores que pueden influir en el rendimiento académico. **(Esta parte fue agregada luego de transformar los datos)**

“Los estudiantes que reciben mayor apoyo parental presentan un GPA promedio más alto que aquellos con menor o nulo apoyo.”

Razonamiento: El acompañamiento familiar favorece la motivación, la constancia y los hábitos de estudio, lo que impacta positivamente en el rendimiento escolar.



Una vez transformada la variable, el gráfico de caja mostró una tendencia ascendente clara: a medida que aumentaba el nivel de *Parental Support*, el GPA promedio también se incrementaba, pasando aproximadamente de **1.5** en los estudiantes sin apoyo (nivel 0) a cerca de **2.2** en los de máximo apoyo (nivel 4).

3. Preprocesamiento y Selección de Variables

Durante esta etapa, se realizó una revisión de la base para asegurar que los datos estuvieran en un formato adecuado para el modelado.

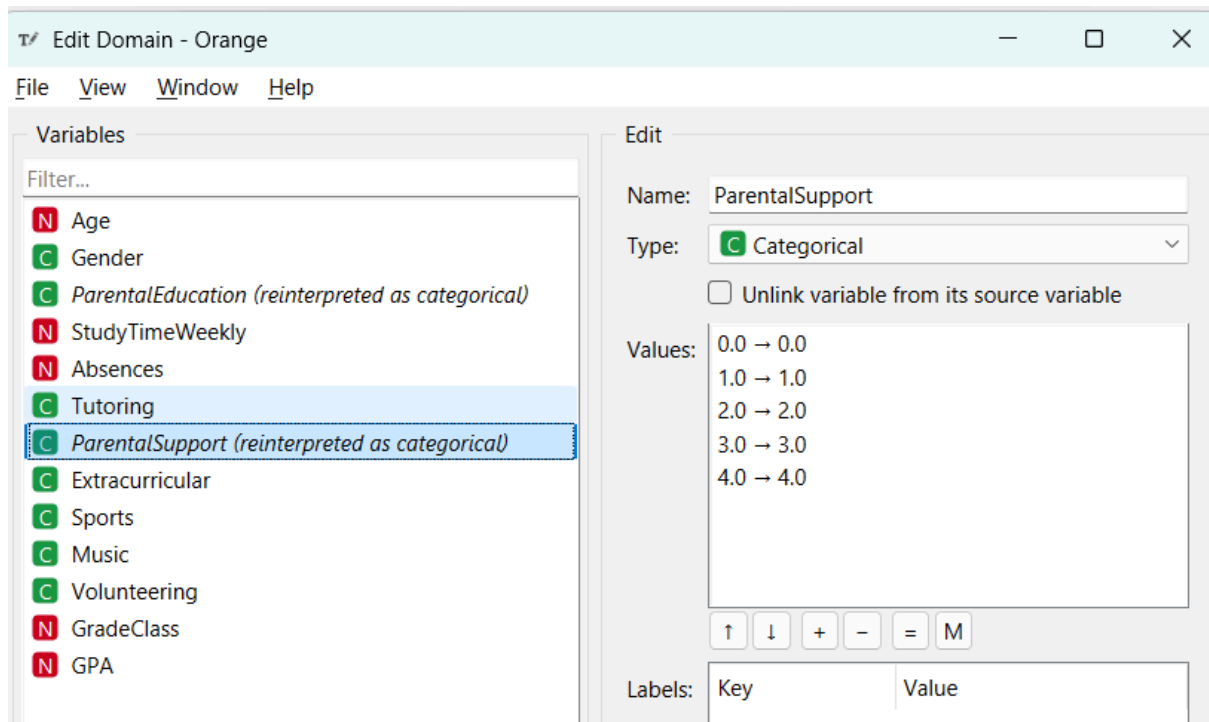
- **Verificación de valores faltantes:**
 - No se encontraron valores faltantes críticos en las variables seleccionadas. Usando el feature statistic vemos que hay missing 0 en cada uno de estos. Por lo tanto no tenemos que completar campos vacíos.
 - El dataset presenta registros completos, lo que facilita el entrenamiento de los modelos sin necesidad de imputación.
- **Detección de valores atípicos:**
 - Se observaron valores extremos en las variables **Absences** y **StudyTimeWeekly** (por ejemplo, estudiantes con 0 o más de 18 horas de estudio semanal, y hasta 29 inasistencias).
 - Dado que estos casos representan comportamientos posibles en contextos reales, se decidió mantenerlos para no perder información útil.
- **Revisión de codificación de variables categóricas:**
 - Variables binarias (**Gender**, **Tutoring**, **Extracurricular**, **Sports**, **Music**, **Volunteering**) ya estaban correctamente codificadas en 0 y 1.
 - No fue necesario aplicar one-hot encoding adicional para estas.

Transformación de Datos

Para preparar los datos para los modelos de machine learning en Orange, se realizaron transformaciones mínimas pero estratégicas:

- **Variable objetivo (**Target**):**
 - Se trabajará con **GPA** como variable continua para tareas de **regresión**.
- **Transformación de variables ordinales:**
 - **ParentalSupport** y **ParentalEducation** fueron convertidas de numéricas a **categóricas ordinales**, ya que representan niveles y no cantidades continuas.

- Esto permite que los modelos interpreten correctamente la naturaleza de estos datos.

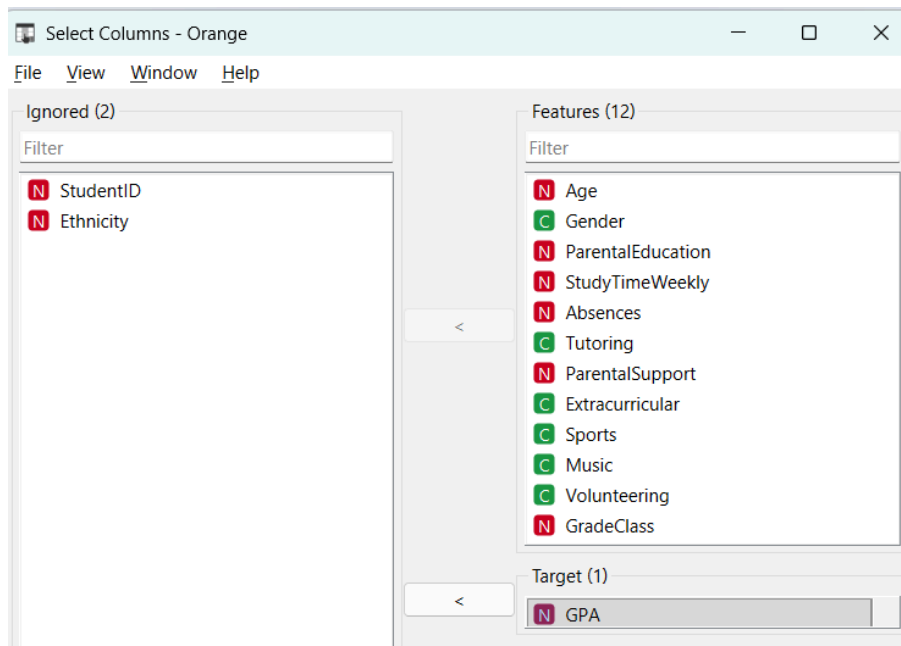


Selección de Variables Relevantes

Basándose en los resultados del análisis exploratorio y en la matriz de correlaciones / ranking de atributos, se definieron las variables más relevantes para predecir el GPA:

- **Variables seleccionadas:**
 - StudyTimeWeekly (relación positiva con GPA).
 - Absences (relación negativa fuerte con GPA).
 - ParentalSupport (nivel ordinal relacionado con rendimiento).
 - ParentalEducation (posible impacto indirecto en desempeño).
 - Gender, Tutoring, Extracurricular, Music (posibles diferencias entre subgrupos).
- **Variables descartadas:**
 - StudentID: es un identificador único sin valor predictivo.

- **Ethnicity**: en esta etapa no se utilizará por falta de contexto interpretativo y relevancia estadística baja.



El análisis con **Rank** permitió reducir el número de variables a las más significativas, optimizando así la construcción del modelo. Esto contribuye a obtener **modelos más simples, interpretables y con mejor rendimiento**.

		#	RReliefF
1	N Absences		0.283
2	C ParentalSupport	5	0.241
3	N GradeClass		0.232
4	N StudyTimeWeekly		0.194
5	N Age		0.193
6	C ParentalEducation	5	0.162
7	C Tutoring	2	0.107
8	C Music	2	0.070
9	C Gender	2	0.065
10	C Extracurricular	2	0.064
11	C Sports	2	0.059
12	C Volunteering	2	0.039

Comentario: La variable *GradeClass* presentó una alta puntuación en el análisis de relevancia debido a su fuerte correlación con el GPA. Sin embargo, no se incluyó en el modelo, ya que representa una versión categorizada del mismo indicador y su incorporación generaría redundancia.

Se optó por utilizar el GPA como variable objetivo principal, manteniendo la independencia entre las variables predictoras y el target.

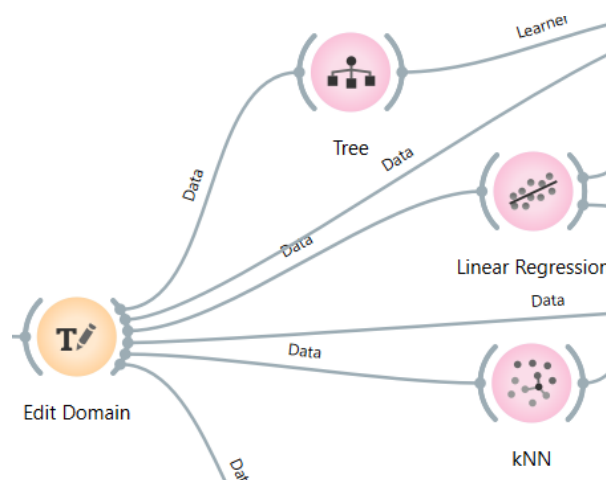
4. Modelado

Selección de modelos a probar

- Se eligieron tres modelos de regresión distintos:
 - Regresión Lineal
 - KNN Regression
 - Tree Regression
- *Propósito*: comparar diferentes algoritmos para ver cuál predice mejor el GPA a partir de las variables explicativas.

Configuración de parámetros

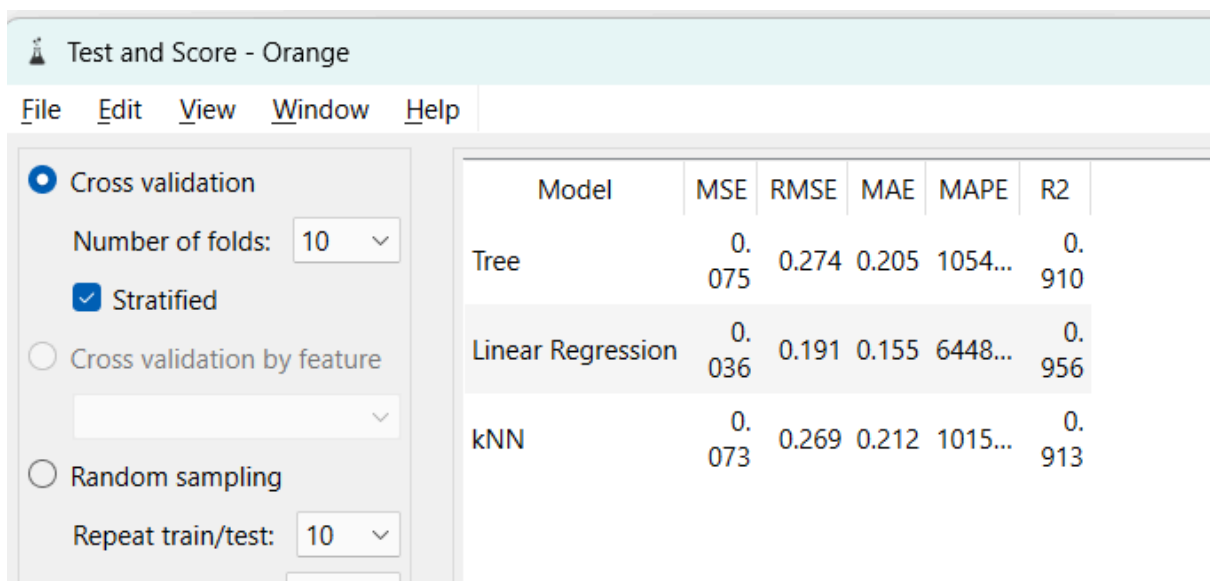
- **Linear Regression**: se dejó con configuración básica, sin regularización (mejor interpretabilidad).
- **KNN**: vecinos = 5, métrica euclidiana, pesos uniformes.
- **Tree**: profundidad máxima limitada y parámetros por defecto para evitar sobreajuste.
- *Propósito*: garantizar que todos los modelos estén correctamente configurados para una comparación justa.



5. Evaluación de Modelos

Evaluación comparativa en Test & Score

- Se conectaron los modelos a *Test & Score* y se utilizó validación cruzada (10 folds).
- Métricas analizadas: **MSE**, **RMSE**, **MAE** y **R²**.
- *Propósito*: medir objetivamente qué modelo se ajusta mejor a los datos.



Model	MSE	RMSE	MAE	MAPE	R2
Tree	0.075	0.274	0.205	1054...	0.910
Linear Regression	0.036	0.191	0.155	6448...	0.956
kNN	0.073	0.269	0.212	1015...	0.913

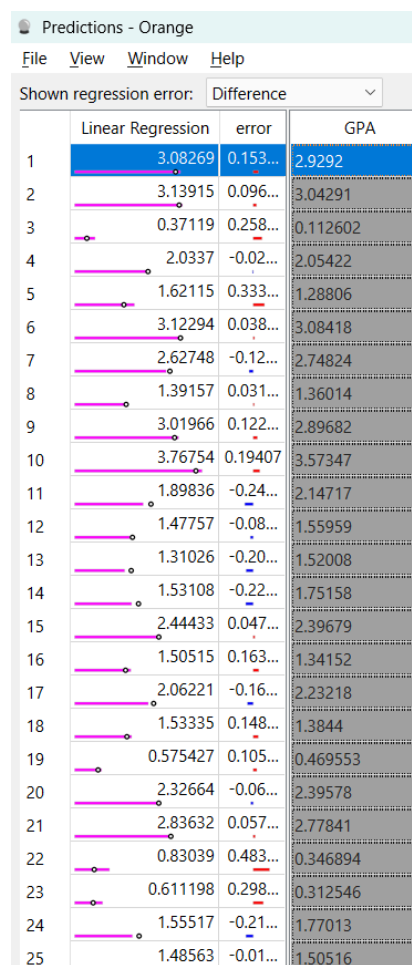
Selección del mejor modelo

- **Regresión Lineal** obtuvo el mejor desempeño con:
 - $R^2 = 0.956$ (muy buen ajuste)
 - $RMSE = 0.191$ (bajo error promedio)
- Tree y KNN tuvieron desempeño más bajo.

- *Conclusión:* se seleccionó **Regresión Lineal** como modelo final por su precisión y facilidad de interpretación.

Generación de predicciones

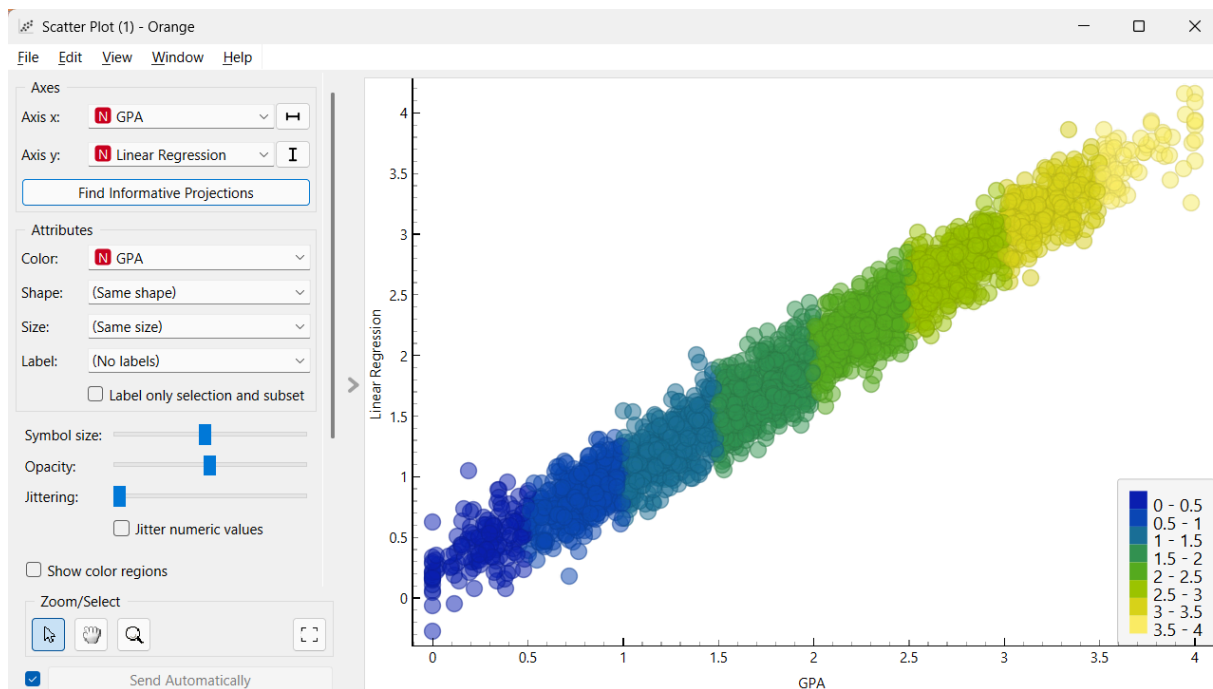
- Se conectaron los datos originales y el modelo lineal al widget **Predictions**.
- Esto permitió generar la columna **Linear Regression** junto al **GPA real**.
- También se puede observar la **diferencia (error)** entre el GPA real y la predicción del modelo.
- **Propósito:** aplicar el modelo sobre todo el conjunto de datos y analizar cómo predice el GPA para cada observación.



	Linear Regression	error	GPA
1	3.08269	0.153...	2.9292
2	3.13915	0.096...	3.04291
3	0.37119	0.258...	0.112602
4	2.0337	-0.02...	2.05422
5	1.62115	0.333...	1.28806
6	3.12294	0.038...	3.08418
7	2.62748	-0.12...	2.74824
8	1.39157	0.031...	1.36014
9	3.01966	0.122...	2.89682
10	3.76754	0.19407	3.57347
11	1.89836	-0.24...	2.14717
12	1.47757	-0.08...	1.55959
13	1.31026	-0.20...	1.52008
14	1.53108	-0.22...	1.75158
15	2.44433	0.047...	2.39679
16	1.50515	0.163...	1.34152
17	2.06221	-0.16...	2.23218
18	1.53335	0.148...	1.3844
19	0.575427	0.105...	0.469553
20	2.32664	-0.06...	2.39578
21	2.83632	0.057...	2.77841
22	0.83039	0.483...	0.346894
23	0.611198	0.298...	0.312546
24	1.55517	-0.21...	1.77013
25	1.48563	-0.01...	1.50516

Visualización de resultados (Scatter Plot)

- Se graficaron los valores reales (GPA) vs. predichos (Linear Regression) en un *Scatter Plot*.
- Los puntos se alinearon muy cerca de la diagonal → el modelo predice con alta exactitud.
- *Propósito: verificar visualmente* la calidad del modelo. Un buen modelo muestra esta alineación fuerte.



6. Interpretación de las Predicciones

Análisis de importancia de variables

- Se revisaron los **coeficientes del modelo lineal** para identificar las variables más influyentes.
- Las más relevantes fueron:

- **Absences** → coeficiente negativo alto → más inasistencias → menor GPA.
 - **StudyTimeWeekly** → coeficiente positivo → más horas de estudio → mayor GPA.
 - **ParentalSupport** y **Tutoring** → efecto positivo moderado.
 - Otras variables (Sports, Music, Gender) → impacto bajo.
- *Propósito*: entender **qué factores explican el rendimiento académico** según el modelo.

Relación con las hipótesis iniciales

- Las hipótesis planteadas al inicio se cumplieron en gran medida:
 - Más horas de estudio → mayor GPA (confirmada)
 - Más inasistencias → menor GPA (confirmada)
 - Tutorías → efecto positivo moderado (parcialmente confirmada)
 - Actividades extracurriculares → poco impacto (no confirmada).
- *Conclusión*: las variables académicas directas (asistencia y estudio) son las más determinantes en el rendimiento.

Conclusión Final

El trabajo permitió aplicar todas las etapas del proceso de análisis de datos en Orange Data Mining, desde la exploración inicial hasta la interpretación final de los resultados. Se partió del objetivo de predecir el rendimiento académico de los estudiantes en función de sus hábitos de estudio, asistencia y apoyo familiar. El análisis exploratorio mostró patrones claros: los estudiantes con más horas de estudio y menos inasistencias obtuvieron un GPA más alto.

Tras el preprocesamiento y la selección de variables relevantes, se entrenaron tres modelos predictivos —Regresión Lineal, KNN y Árbol de Decisión— evaluados mediante validación cruzada con 10 folds. La Regresión Lineal obtuvo el mejor desempeño ($R^2 = 0.956$), mostrando un ajuste excelente entre valores reales y predichos.

Las predicciones confirmaron las hipótesis iniciales: el tiempo de estudio y la asistencia influyen directamente en el GPA, mientras que el apoyo parental tiene un impacto positivo sostenido, siendo de las variables más relacionadas con el problema planteado. En síntesis,

el rendimiento académico depende principalmente de la dedicación, la constancia y el acompañamiento familiar, factores que deberían guiar futuras estrategias educativas.