Trabajo Práctico – Big Data

Análisis y Modelado de Datos con Orange Data Mining

Grupo: Diving Insight (Sist, Denevi, Ritrovato, Martin Botter y Beumont)

Dataset: *Students Performance Dataset (Rabie El Kharoua, Kaggle)*

Variable objetivo: *GradeClass*

1- Definición del problema

En este trabajo buscamos predecir el nivel de rendimiento académico de los estudiantes (GradeClass) a partir de variables relacionadas con sus hábitos de estudio, características personales y apoyo familiar.

Seleccionamos este target porque la variable *GradeClass* está definida explícitamente en la descripción del conjunto de datos y permite clasificar a los estudiantes según sus calificaciones académicas.

El objetivo es identificar los factores que más influyen en el rendimiento escolar y construir un modelo predictivo que clasifique a los alumnos en diferentes categorías de desempeño (A, B, C, D, F). Esto resulta relevante para el ámbito educativo, ya que puede orientar estrategias de apoyo a estudiantes con riesgo de bajo rendimiento.

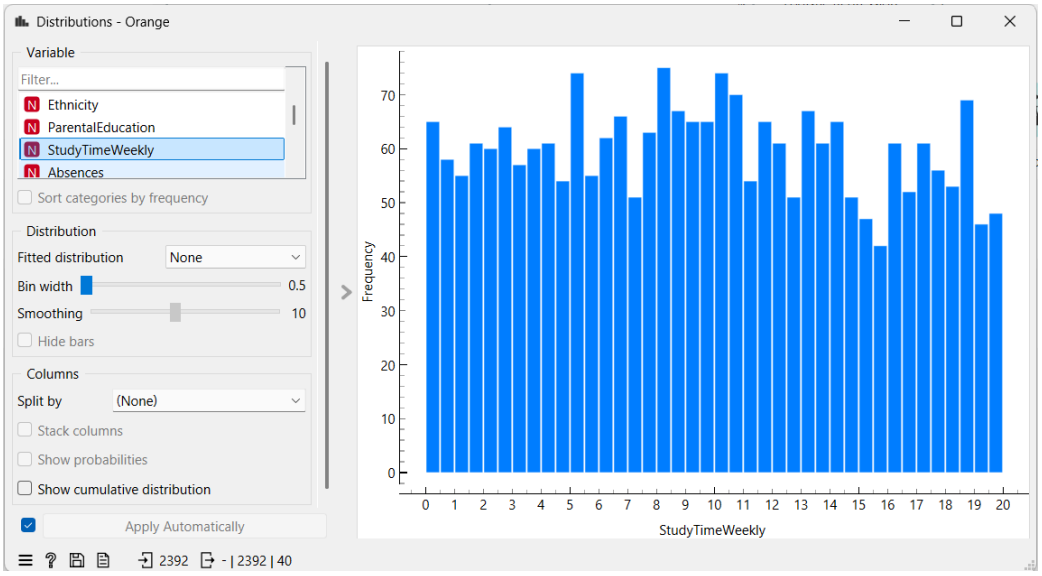
La escuela que analizamos atraviesa un contexto de recursos económicos limitados, lo que le impide ofrecer programas de apoyo generalizados a todos los alumnos.

Por este motivo, se busca utilizar un modelo de inteligencia artificial predictiva que ayude a asignar los recursos educativos de manera eficiente, enfocando el acompañamiento solo en los estudiantes con mayor riesgo de bajo rendimiento.

2- Análisis exploratorio de datos (EDA)

Realizamos un análisis exploratorio inicial utilizando los módulos Distributions y Feature Statistics de Orange, con el objetivo de comprender la distribución de las variables y detectar posibles relaciones con el rendimiento académico.

Se observó que las variables relacionadas con hábitos de estudio (StudyTimeWeekly) y ausencias (Absences) presentan una variabilidad importante entre los alumnos, mientras que la variable GPA tiene una relación directa con *GradeClass*.
Por este motivo, decidimos excluir GPA del modelo, ya que su inclusión podría sesgar el aprendizaje del algoritmo (al estar directamente vinculada con la clase objetivo).



3- Preprocesamiento de datos

Antes del modelado, se realizó una selección de variables mediante el nodo Select Columns, conservando aquellas más relevantes para el análisis.

Variables utilizadas:

Features:

- Gender
- Ethnicity
- ParentalEducation
- StudyTimeWeekly
- Absences
- Tutoring
- ParentalSupport
- Extracurricular
- Sports
- Music
- Volunteering

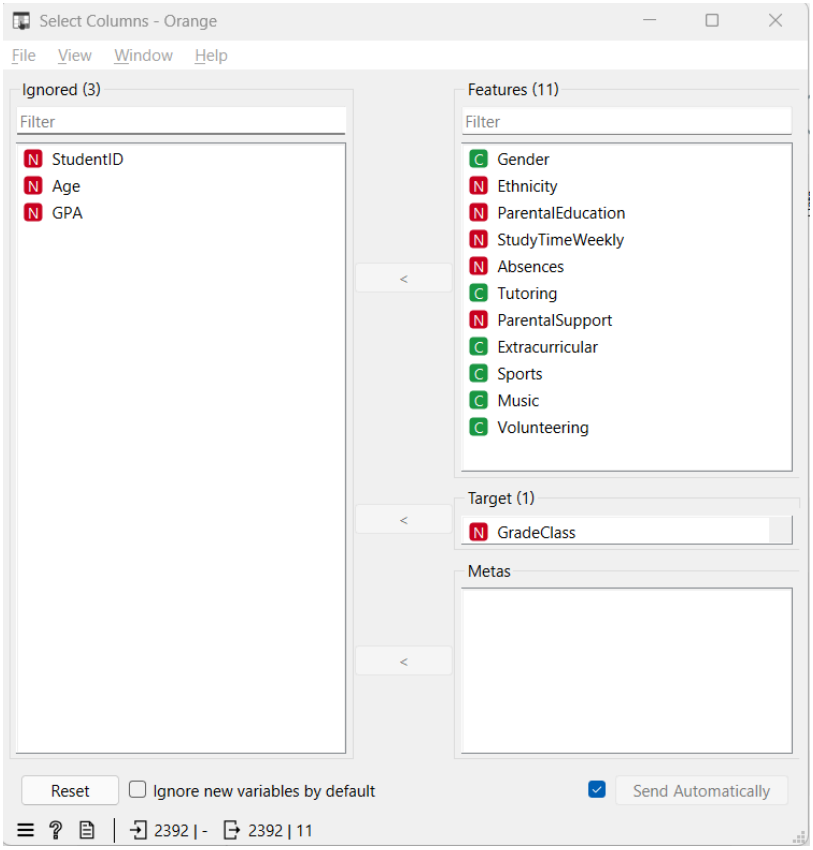
Target:

- GradeClass

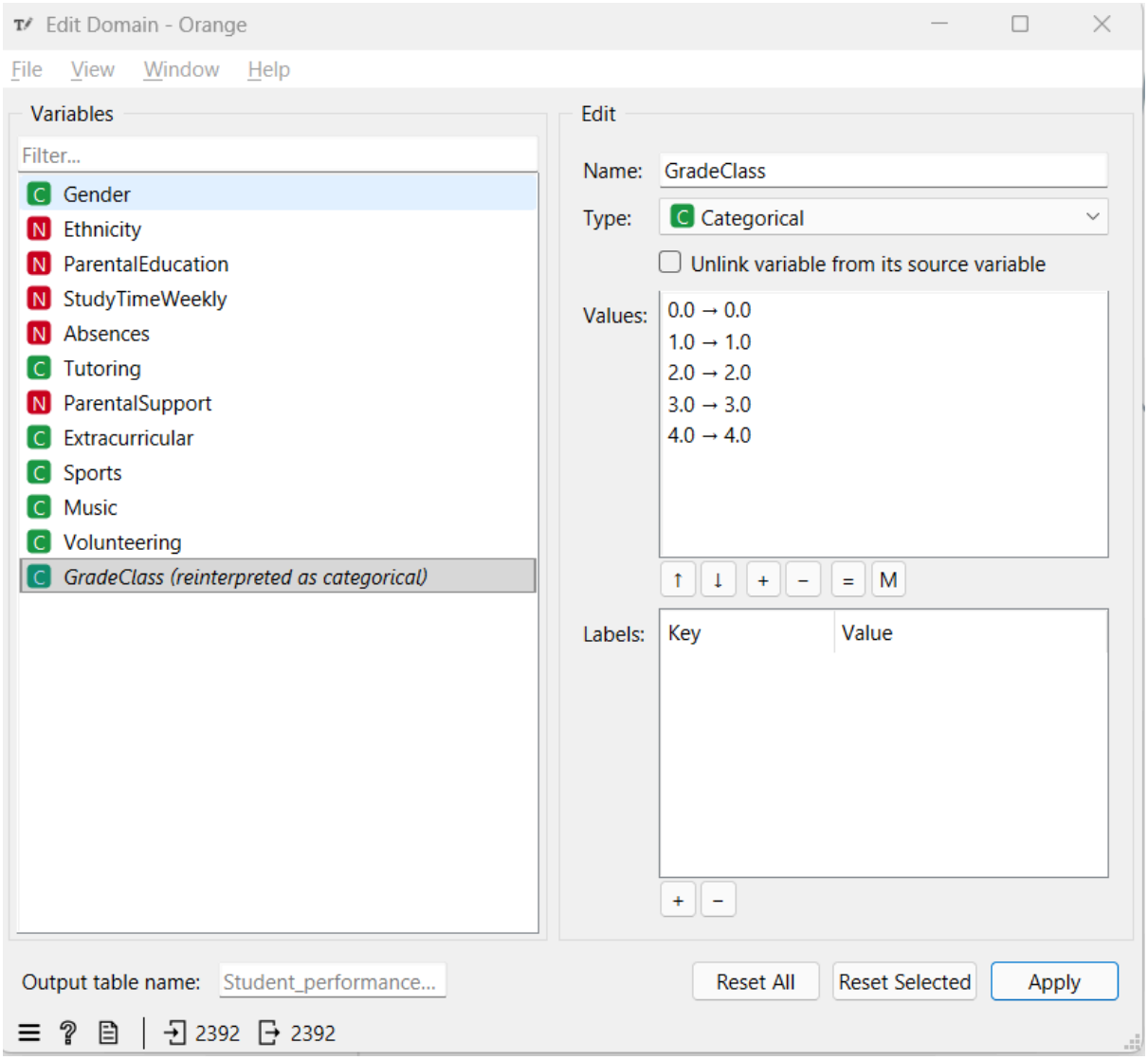
Eliminadas:

- StudentID (no aporta información predictiva)

- Age (variación mínima)
- GPA (define directamente la clase objetivo)



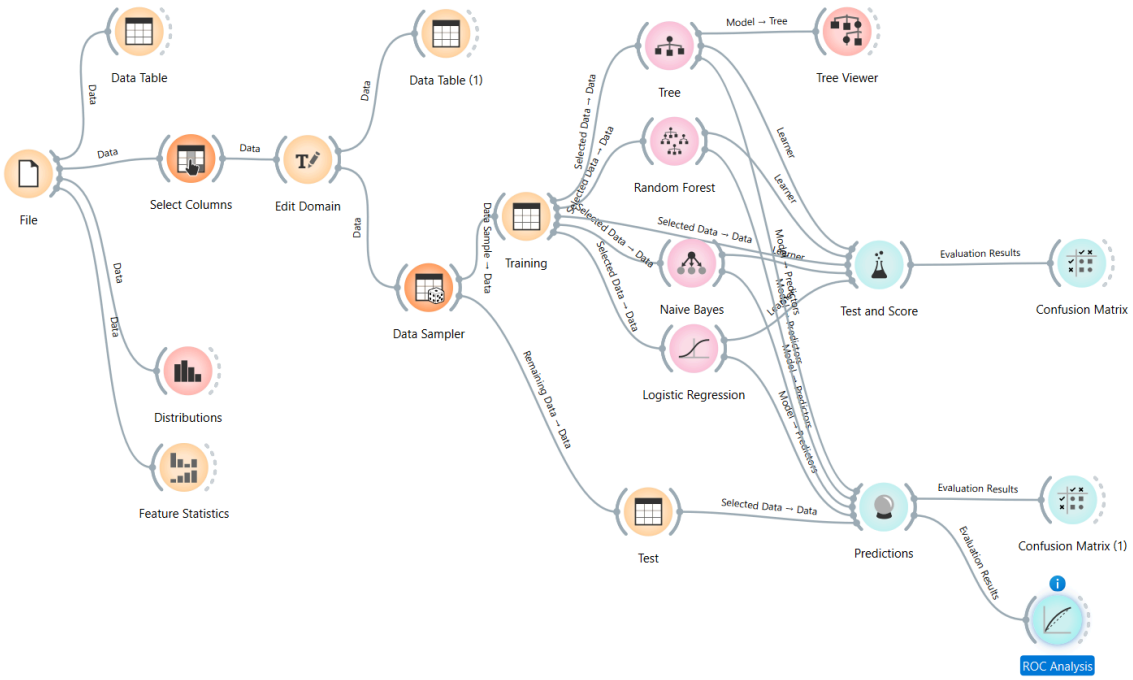
Ademas cambiamos la variable GradeClass de numerica a categorica.



4- Modelado

Probamos cuatro algoritmos de aprendizaje automático incluidos en Orange:

- Logistic Regression
- Naive Bayes
- Random Forest
- Decision Tree



5- Evaluación de los modelos

Los modelos se evaluaron utilizando validación cruzada y métricas de desempeño (CA, F1, Precision, Recall, MCC y AUC) mediante el nodo Test & Score.

Model	AUC	CA	F1	Prec	Recall	MCC
Logistic Regression	0.901	0.728	0.703	0.686	0.728	0.577
Random Forest	0.862	0.688	0.674	0.669	0.688	0.517
Naive Bayes	0.860	0.658	0.653	0.657	0.658	0.482
Tree	0.744	0.666	0.668	0.671	0.666	0.499

El mejor desempeño lo obtuvo el modelo Logistic Regression, con una precisión (CA) de 0.728 y un AUC de 0.901, lo que indica una excelente capacidad discriminativa. Consideramos que este modelo logra el mejor equilibrio entre simplicidad y rendimiento, superando a los árboles y modelos probabilísticos.

En nuestro caso, la escuela cuenta con recursos muy limitados para brindar apoyo académico, por lo que necesita asegurar que los esfuerzos se destinen a quienes realmente lo requieren.

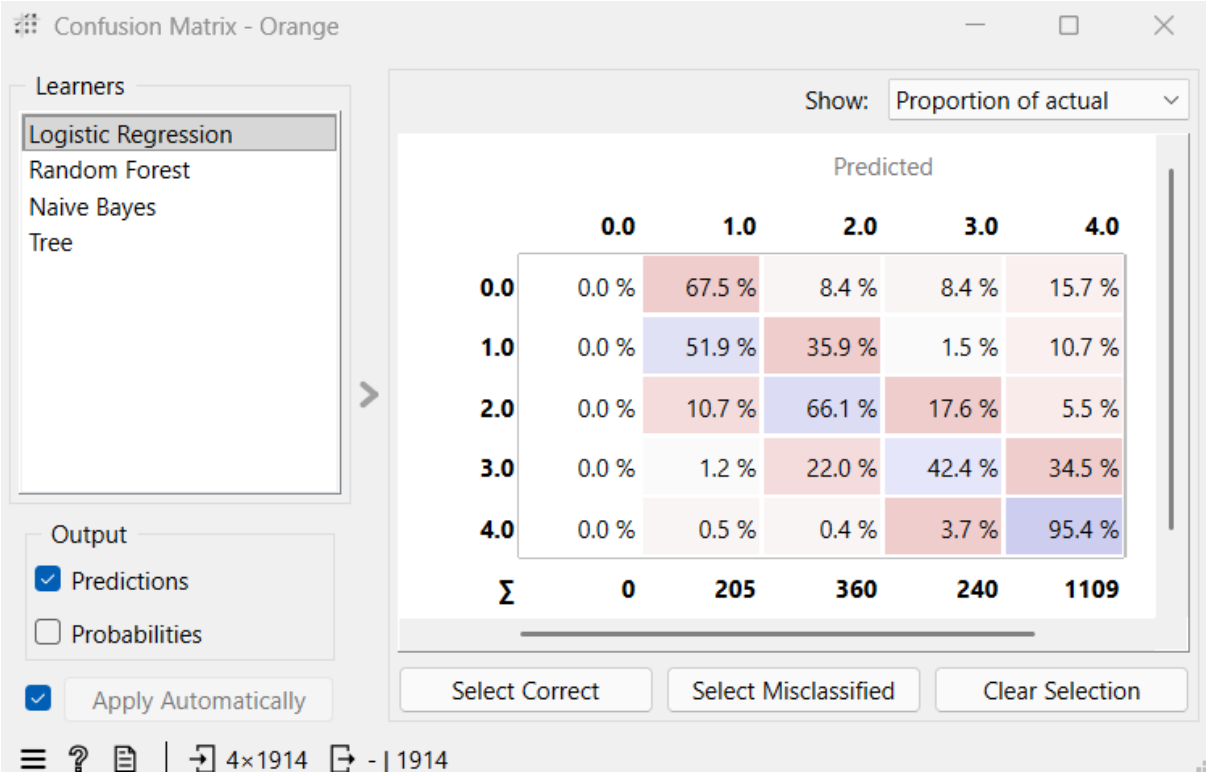
Cada hora de tutoría o acompañamiento debe asignarse con criterio, para no invertir en alumnos que igual obtendrán buenos resultados sin ayuda.

Por eso, decidimos priorizar el Recall como métrica principal.

Al enfocarnos en esta métrica, buscamos maximizar la detección de los estudiantes que efectivamente presentan riesgo académico, garantizando que el sistema no deje sin asistencia a quienes más la necesitan.

Aunque esto pueda implicar un margen pequeño de error (ayudar a algún alumno que hubiera aprobado igual), el objetivo principal es no malgastar recursos en intervenciones erróneas ni dejar sin apoyo a los casos reales de bajo rendimiento.

6- Matriz de confusión



En la imagen se observa la matriz de confusión correspondiente al modelo de *Regresión Logística*, expresada como proporción respecto de las clases reales (*Proportion of actual*). El modelo muestra un comportamiento diferenciado según el nivel de rendimiento académico. Los estudiantes con calificaciones más bajas (*F*) fueron clasificados correctamente en un **95,4 % de los casos**, lo que evidencia una excelente capacidad para identificar el bajo desempeño. Sin embargo, los alumnos con calificaciones altas (*A*) fueron subestimados en su mayoría, siendo clasificados como *B* en un **67,5 %**, lo que refleja una menor precisión en los niveles superiores. Las clases intermedias (*B*, *C* y *D*) presentan una mayor confusión entre sí, lo cual es esperable dado que sus límites son menos definidos. En general, el modelo logra distinguir con claridad los extremos de rendimiento (alto y bajo), aunque le resulta más difícil separar categorías adyacentes.

7- Conclusiones

A partir de este trabajo pudimos:

- Comprender la relación entre las variables académicas y demográficas de los estudiantes.
- Construir un pipeline completo en Orange para el análisis y modelado predictivo.
- Identificar los factores más influyentes en el rendimiento académico.
- Determinar que el modelo de Regresión Logística ofrece la mejor capacidad