Estudio de performance de estudiantes de escuelas secundarias.

Agenda

1. Contexto y Audiencia.
2. Hipótesis y preguntas de Interés.
3. Descripción de metadatos.
4. Análisis Exploratorio.
5. Selección de modelos.
6. Optimización.
7. Insights y Recomendaciones.

Contexto y Audiencia

Se desea interpretar las variables que influyen en el desempeño académico de un grupo de estudiantes de escuelas secundarias. El estudio está dirigido a entidades gubernamentales como el Ministerio de Educación y a autoridades de establecimientos educativos.

Hipótesis y Preguntas

- Qué incidencia y peso tienen las variables actitudinales como hábitos de estudio, horas de estudio, ausencias, tutorías, participación de los padres y actividades extracurriculares?
- Se relaciona el desempeño con factores demográficos? Estos pueden incluir educación de los padres, minorías sociales que indiquen si se necesita investigar más en profundidad el acceso a los recursos educativos
- Investigar si hay diferencias significativas en las notas separando por grupos, por ejemplo género.
- Sentar las bases para elaborar modelos de predicción de desempeño.

Descripción de los metadatos

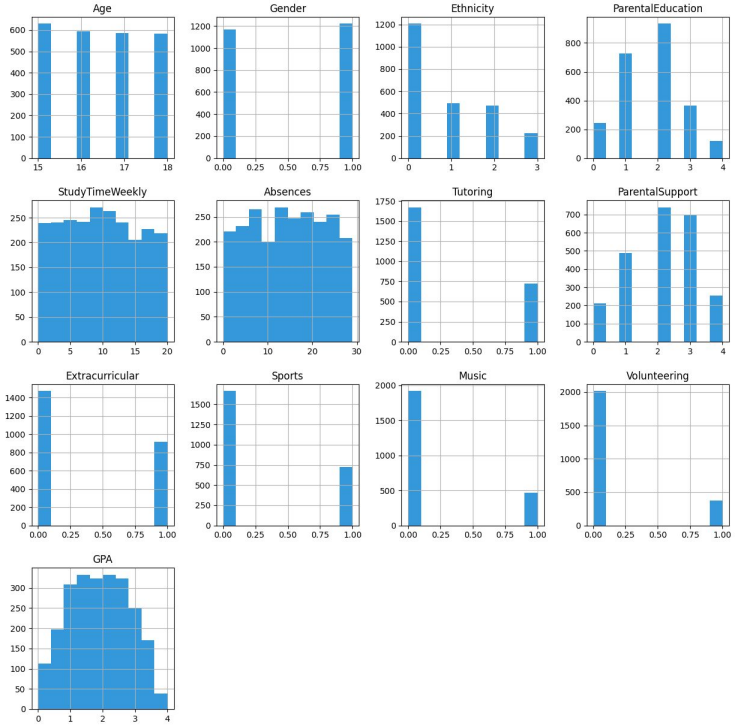


Fig. 1 Descripción del dataset

- 2392 estudiantes
- 13 características de interés
- Variable objetivo: GPA
- Distribución de edades homogénea (15, 16, 17 y 18 años)
- Balance de género.
- Predominio de estudiantes caucásicos.
- Asimetría en los conteos por si o por no de las actividades extra-áulicas (Extracurriculares, Deportes, Música, Voluntariados.)

CODERHOUSE

Análisis Exploratorio

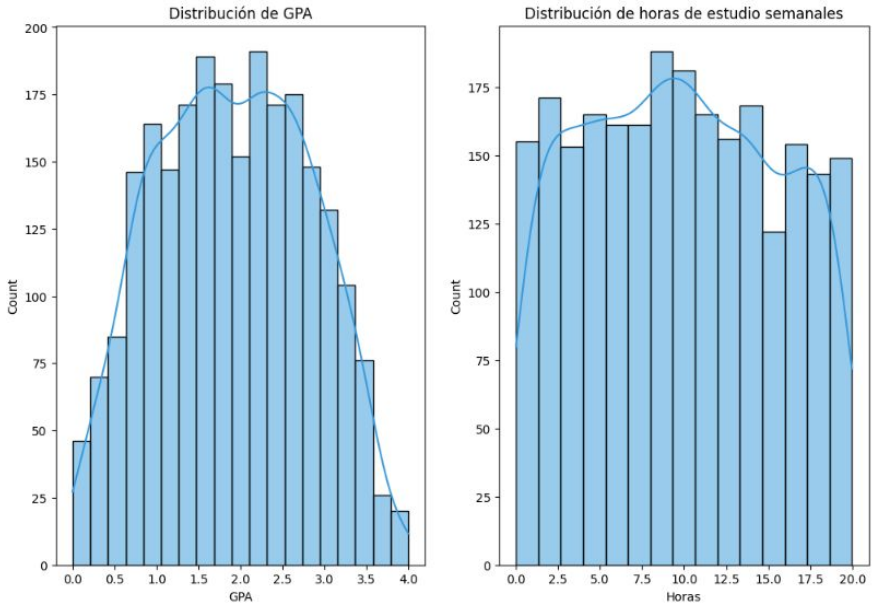


Fig. 2 Distribuciones de GPA y Horas de estudio semanales

- La distribución de notas GPA presenta una forma acampanada que asemeja a una distribución normal recortada por el rango entre 0 y 4.
- La cantidad de horas asignadas a estudio varía entre 0 y 20.

Análisis Exploratorio

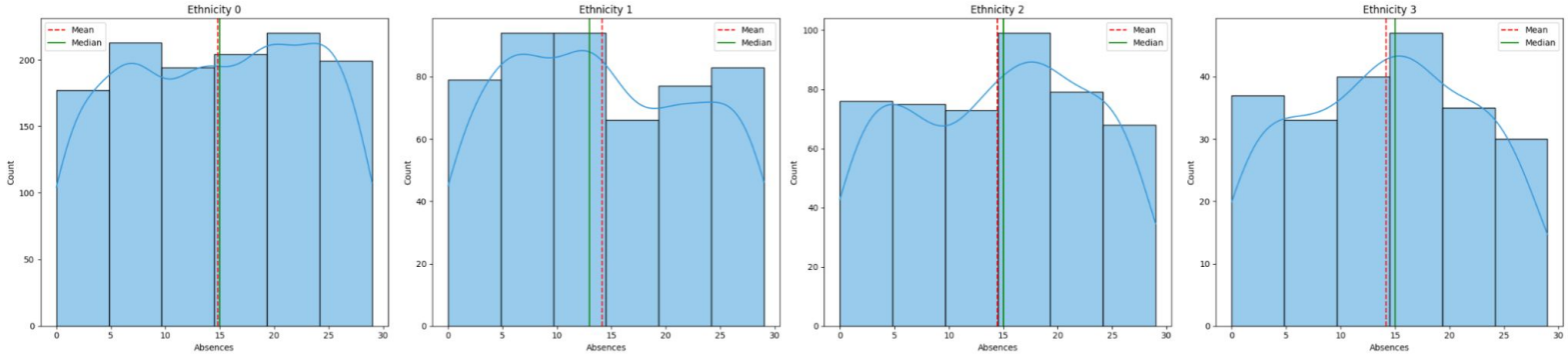


Fig. 3 Ausencias por etnias.

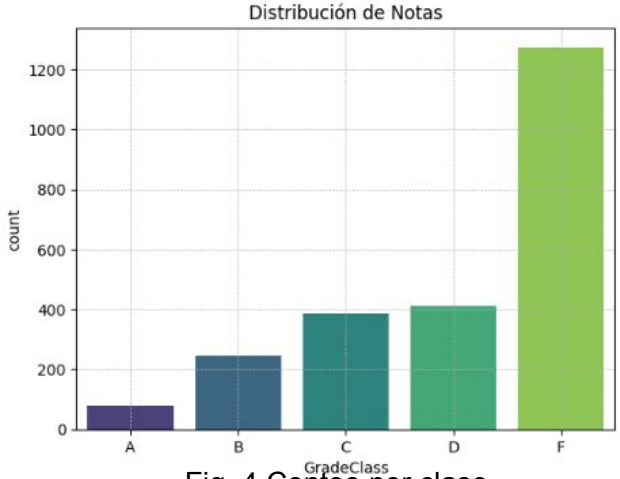


Fig. 4 Conteo por clase

- Las ausencias no tienen relación evidente con las etnias (Fig. 3)
- Se evidencia una proporción mucho más grande ~50% de desaprobados (nota F)
- Son mucho menos frecuentes en el otro extremo las “A”. (Fig. 4)

CODERHOUSE

Análisis Exploratorio

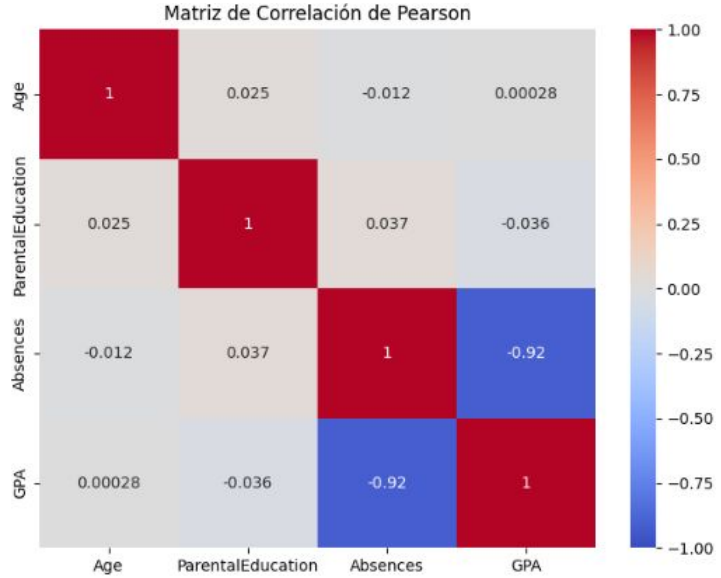


Fig. 5 Matriz de Correlación

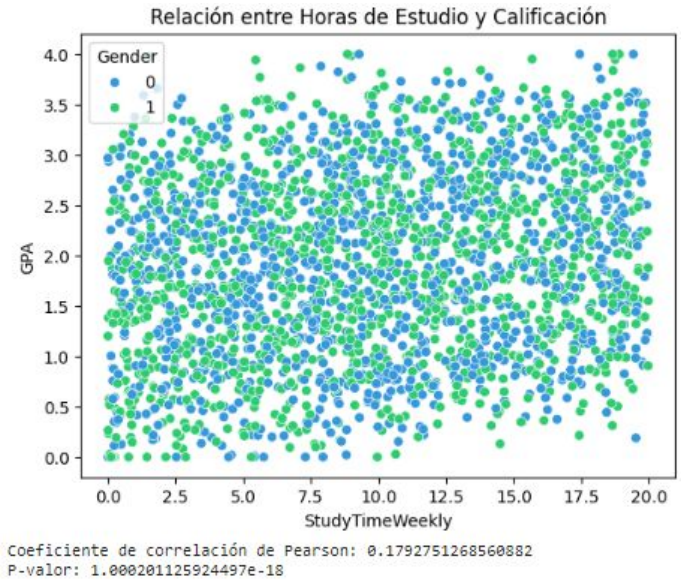


Fig. 6 GPA vs Horas

Se identifica una correlación negativa fuerte entre ausentismo y performance (Fig. 5). Hay también una correlación positiva débil entre horas de estudio y GPA. (Fig. 6) y lo mismo pasa con los estudiantes que poseen tutorías. El resto de las variables, aún asumiendo la posibilidad de relaciones no lineales, no impresionan tener relación con el desempeño.

CODERHOUSE

Selección de modelos

Model	Accuracy	Balanced Accuracy	ROC AUC	F1 Score	Time Taken
LGBMClassifier	0.95	0.94	0.94	0.95	0.08
CalibratedClassifierCV	0.94	0.94	0.94	0.94	0.04
LogisticRegression	0.94	0.94	0.94	0.94	0.03
LinearSVC	0.94	0.94	0.94	0.94	0.03
QuadraticDiscriminantAnalysis	0.94	0.93	0.93	0.94	0.02
AdaBoostClassifier	0.94	0.93	0.93	0.94	0.13
BaggingClassifier	0.94	0.93	0.93	0.94	0.06
SVC	0.93	0.93	0.93	0.93	0.14
RandomForestClassifier	0.93	0.93	0.93	0.93	0.33
SGDClassifier	0.93	0.93	0.93	0.93	0.03
XGBClassifier	0.93	0.93	0.93	0.93	0.18
RidgeClassifierCV	0.93	0.93	0.93	0.93	0.03
RidgeClassifier	0.93	0.93	0.93	0.93	0.02
LinearDiscriminantAnalysis	0.93	0.93	0.93	0.93	0.02

Fig. 7 Chequeo de modelos

Se testearon y compararon diferentes modelos, LogisticRegression, SVC, RandomForestClassifier y XGBClassifier. Considerando tiempos y precisión balanceada, se seleccionó LogisticRegression.

F1 scores, revisado por separado

```
F1-score para Random Forest: 0.9285714285714286
F1-score para XGBoost: 0.9251700680272109
F1-score para SVC: 0.9227272727272727
F1-score para LogisticRegression: 0.9330357142857143
```

Optimización

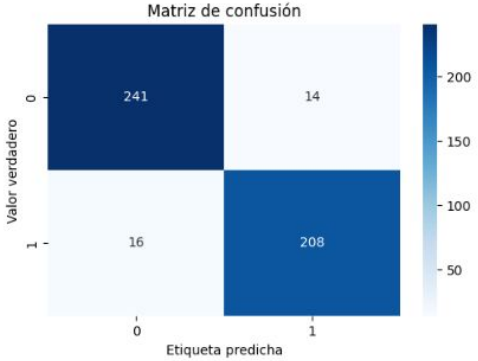
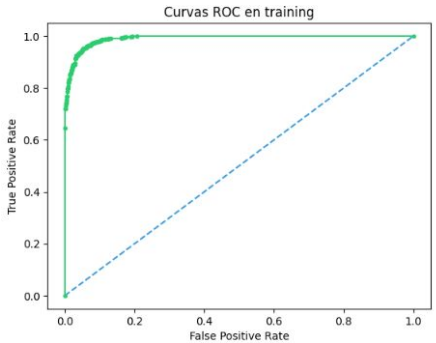


Fig. 8 Confusion matrix



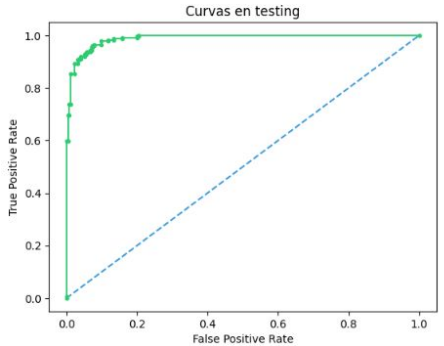


Fig. 9 ROC Curves and AUC

Se testearon y compararon diferentes modelos, LogisticRegression, SVC, RandomForestClassifier y XGBClassifier. Considerando tiempos y precisión balanceada, se seleccionó LogisticRegression.

F1 scores, revisado por separado

```
F1-score para Random Forest: 0.9285714285714286
F1-score para XGBoost: 0.9251700680272109
F1-score para SVC: 0.9227272727272727
F1-score para LogisticRegression: 0.9330357142857143
```

La Puntuación media de F1 en validación cruzada: 0.94

CODERHOUSE

Insights y conclusiones

Se realizó el estudio del dataset satisfactoriamente, analizando las distintas variables teniendo en cuenta como variable objetivo la calificación.

💬 Hallazgos principales:

- Las ausencias de los estudiantes son la principal causa en la pérdida de performance académicas en este grupo de escuelas secundarias.
- Los alumnos que cuentan con tutorías, mejoran levemente el desempeño.
- Hay tendencia en alumnos que estudian más por semana a que les vaya mejor.
- No se encontraron factores demográficos que inciden en la nota, diferencias por género, etnia, etc.
- El ausentismo no está relacionado con etnias ni grupos minoritarios, tampoco se relaciona con el género.
- Los modelos de predicción evaluados arrojaron buena performance, identifican correctamente las clases y tienen buena capacidad de generalización.
- La validación cruzada se utilizó para verificar la robustez del modelo, los resultados muestran buen equilibrio entre precisión y sensibilidad. Las curvas ROC y AUC muestran un rendimiento estable y elevado, también sugiere que el modelo distingue correctamente entre clases.
- Enfoque que permitió disminuir el riesgo de un ajuste excesivo al conjunto de entrenamiento.