Examenes estudiantes

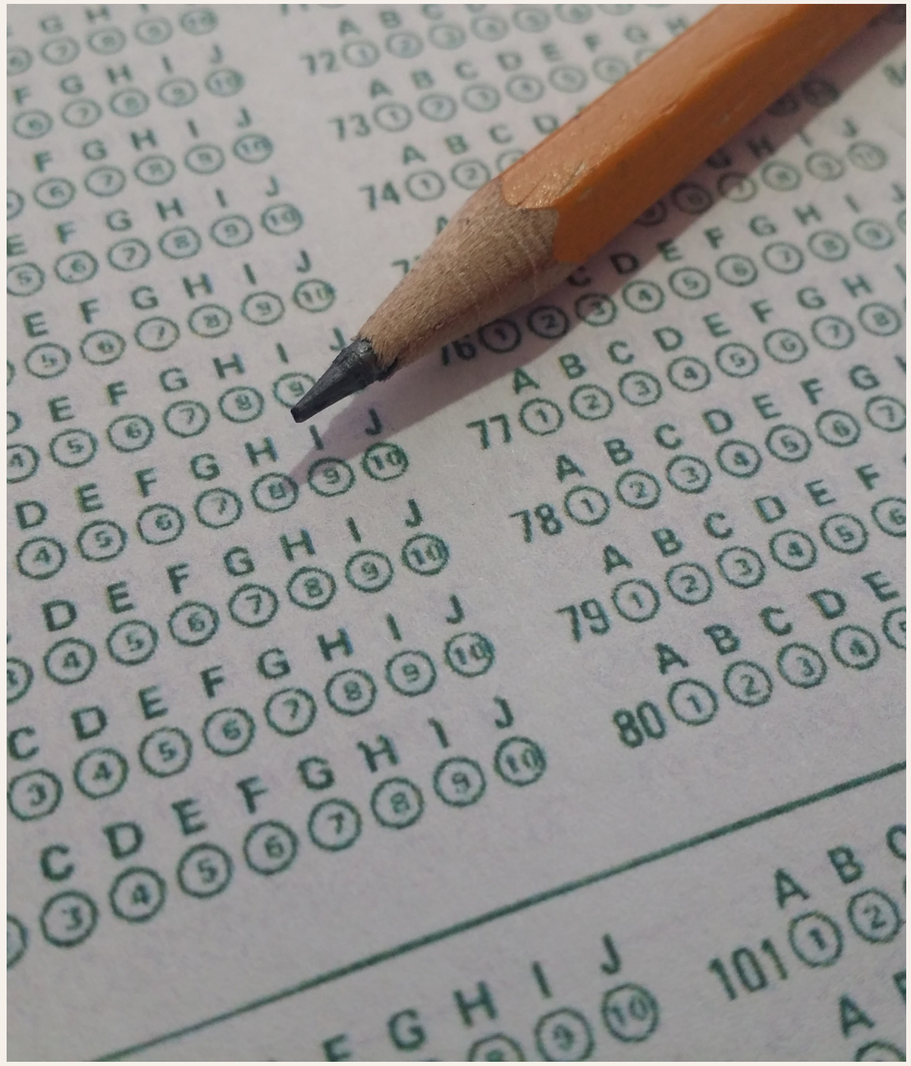
Jeremias Iturriza

Agustin Rebechi

Formulacion del problema

Nuesto objetivo es predecir si un alumno aprobará o no en función de diversas características. Esto ayudaria a identificar estudiantes en riesgo y brindar apoyo necesario, proporcionando una herramienta útil para ayudar a profesores e instituciones educativas en toma de decisiones que mejoren el rendimiento académico.

Es por eso que utilizaremos tecnicas de machine learning para este problema de clasificacion



Presentacion del dataset

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 30641 entries, 0 to 30640
Data columns (total 14 columns):
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   Gender                30641 non-null  object
 1   EthnicGroup           28801 non-null  object
 2   ParentEduc            28796 non-null  object
 3   LunchType             30641 non-null  object
 4   TestPrep              28811 non-null  object
 5   ParentMaritalStatus   29451 non-null  object
 6   PracticeSport         30010 non-null  object
 7   IsFirstChild          29737 non-null  object
 8   NrSiblings            29069 non-null  float64
 9   TransportMeans        27507 non-null  object
 10  WklyStudyHours        29686 non-null  object
 11  MathScore             30641 non-null  int64
 12  ReadingScore          30641 non-null  int64
 13  WritingScore          30641 non-null  int64
dtypes: float64(1), int64(3), object(10)
memory usage: 3.3+ MB
```

Fuente: https://www.kaggle.com/datasets/desalegngeb/students-exam-scores

- 30641 filas
- 14 columnas

Variables categóricas:

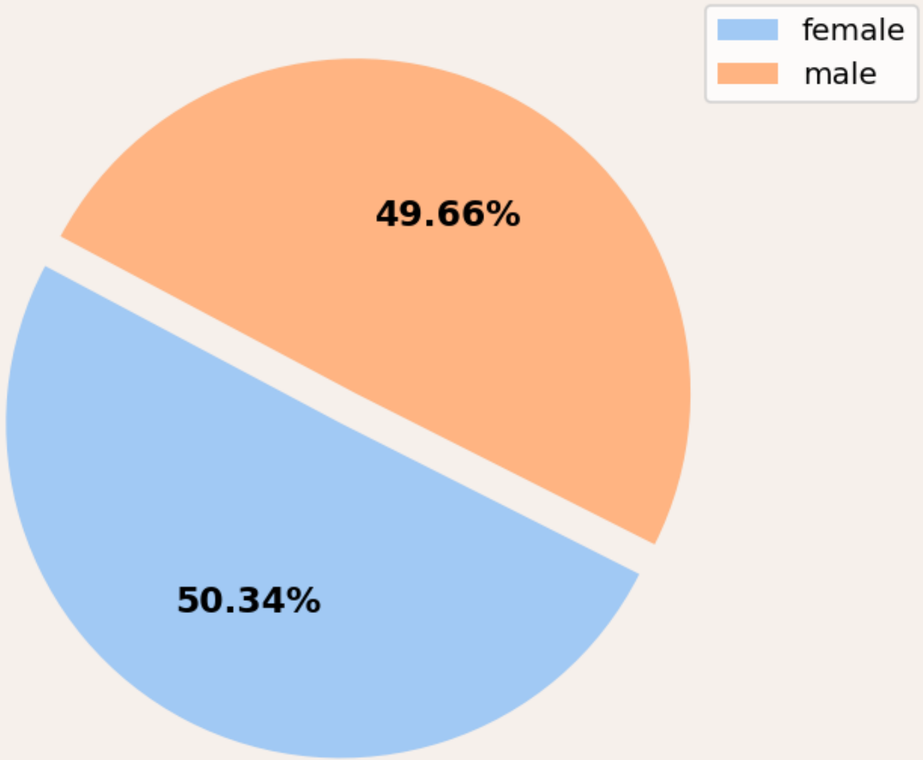
Gender, EthnicGroup, ParentEducation, LunchType, TestPrep, ParentMaritalStatus, PractiveSport, TransportMeans, WklyStudyHours

Variables numéricas:

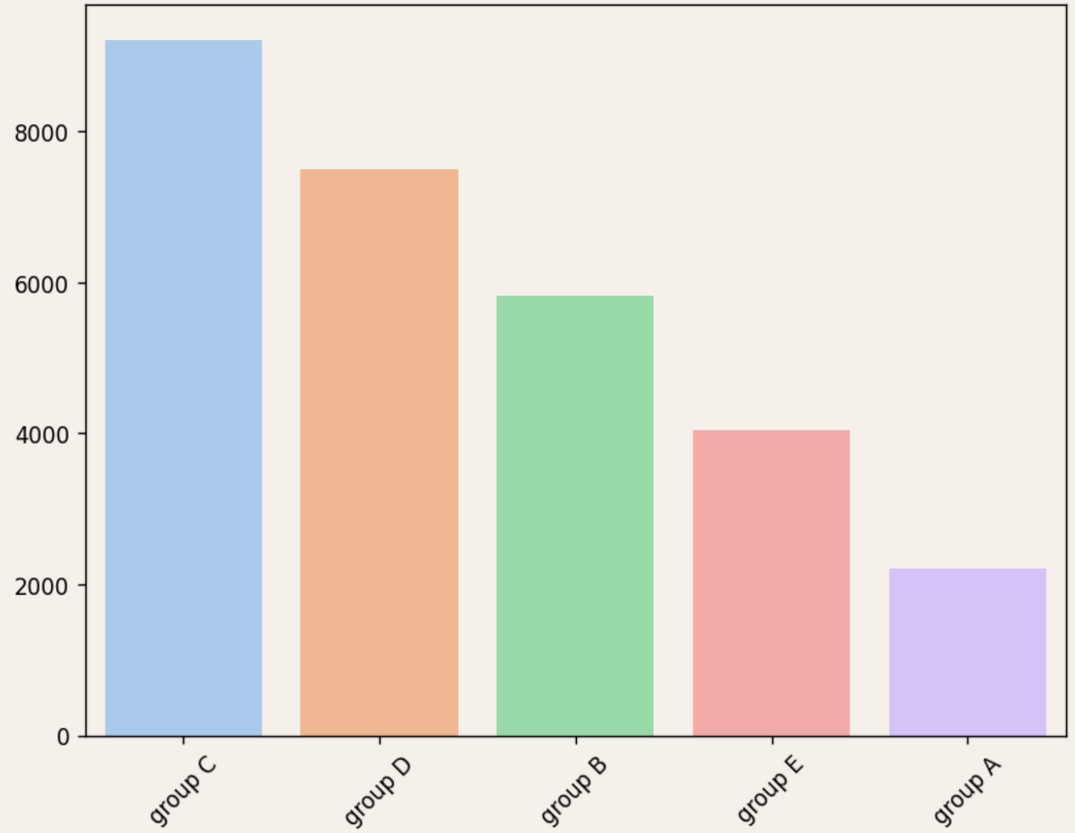
NrSilbings, MathScore, ReadingScore, WritingScore

Analisis exploratorio

Genero

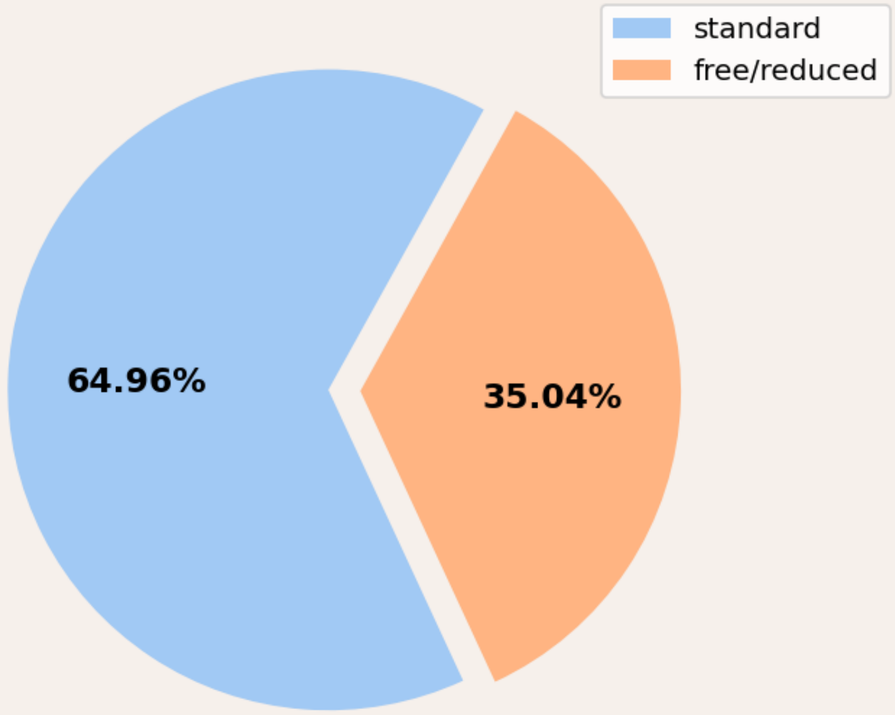


Grupo etnico

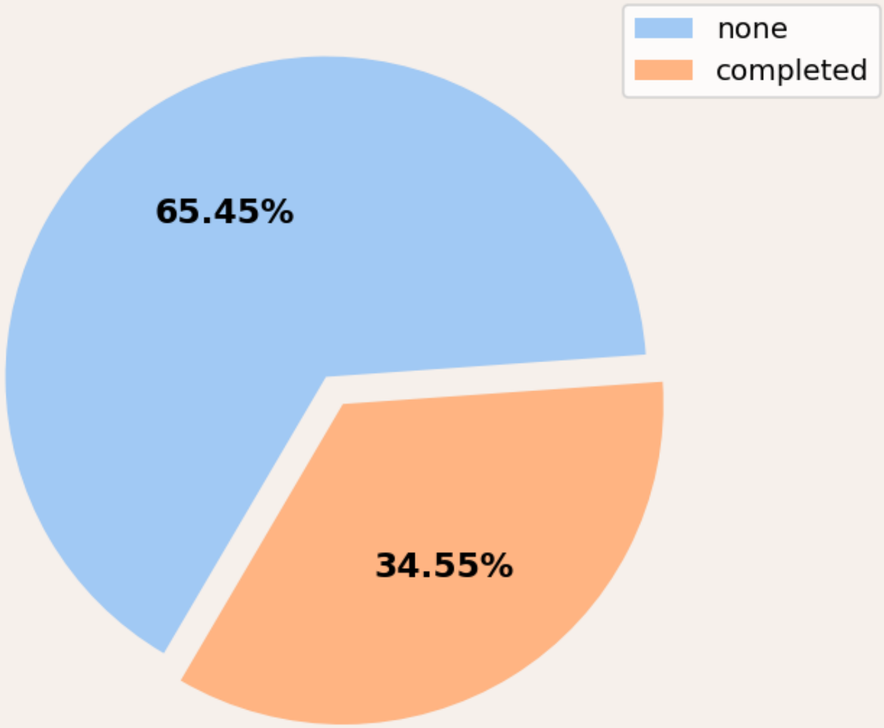


Analisis exploratorio

Almuerzo

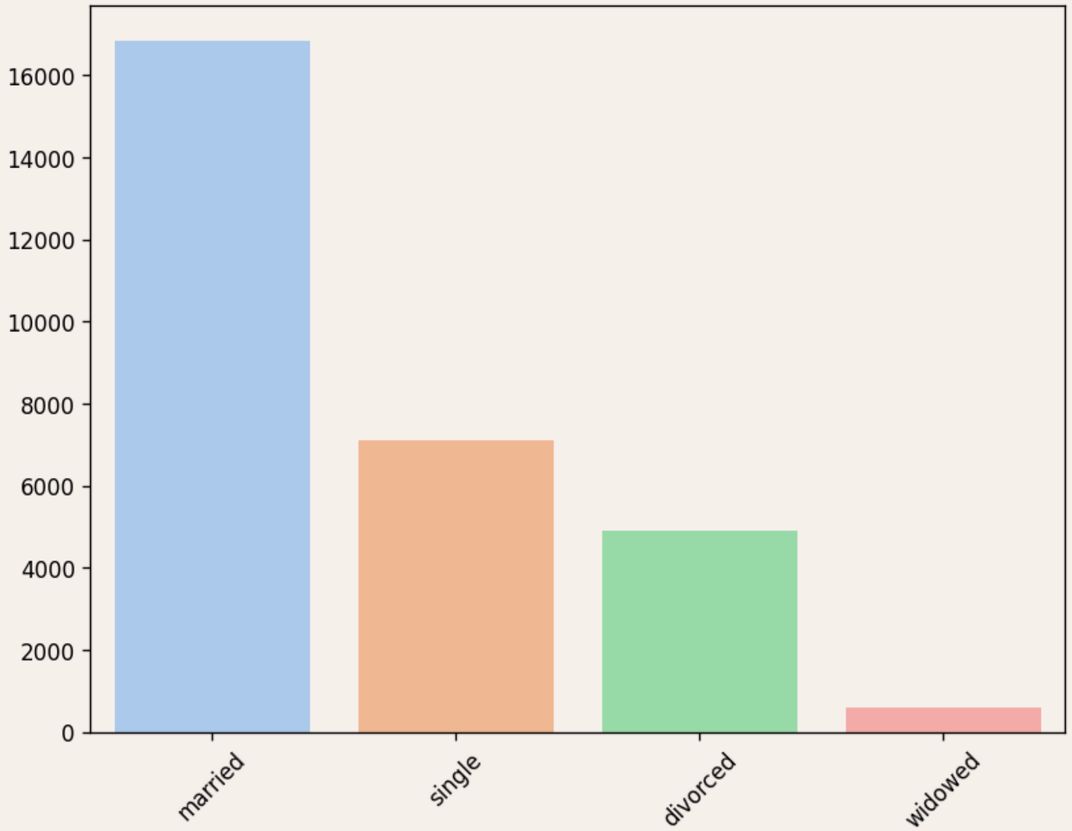


Curso de preparacion

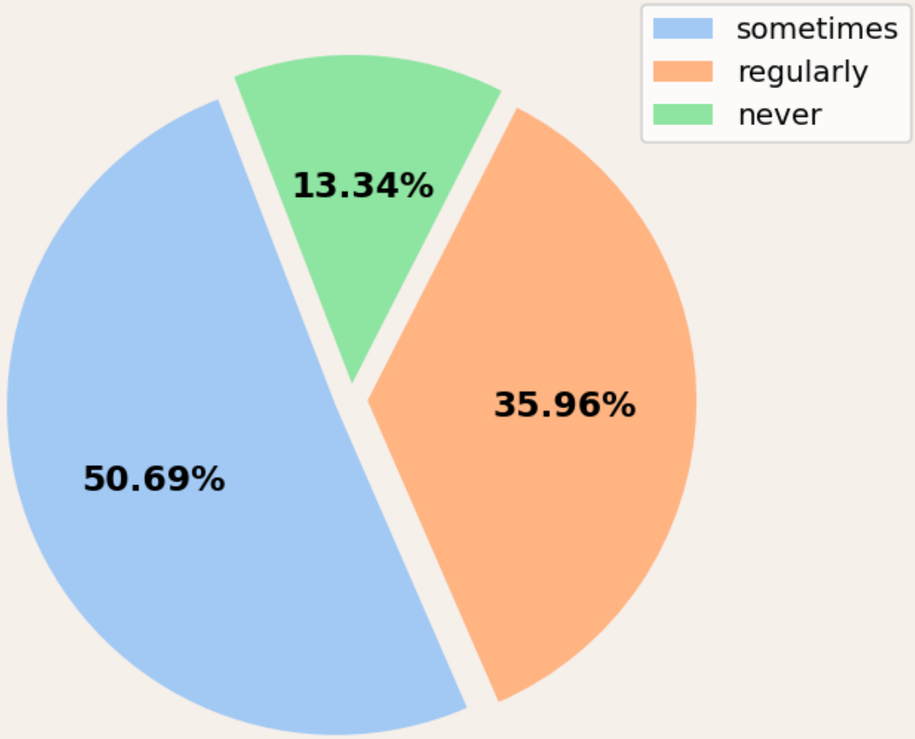


Analisis exploratorio

Estado civil de los padres

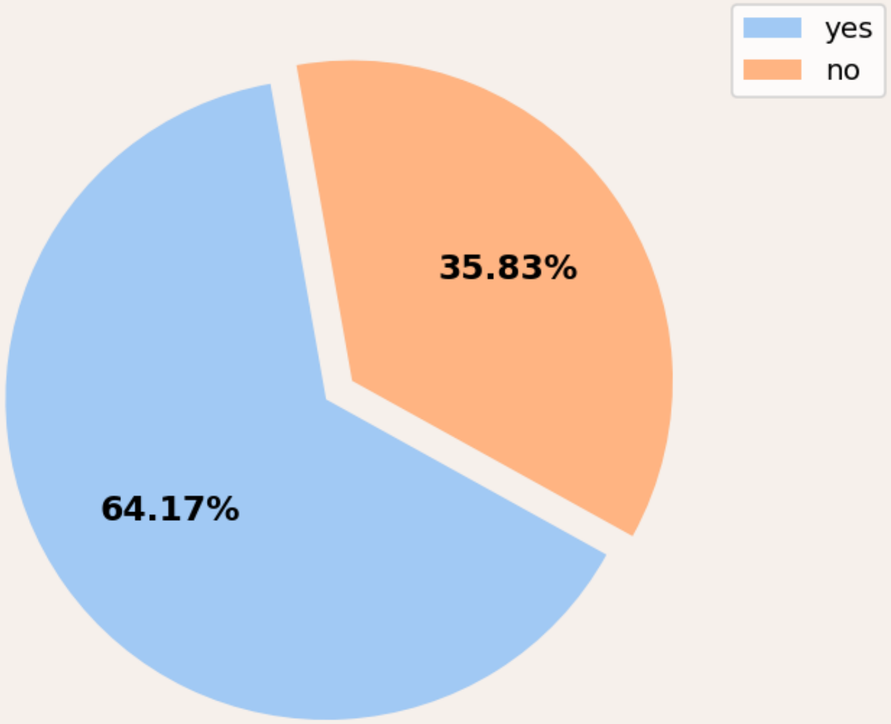


Deporte

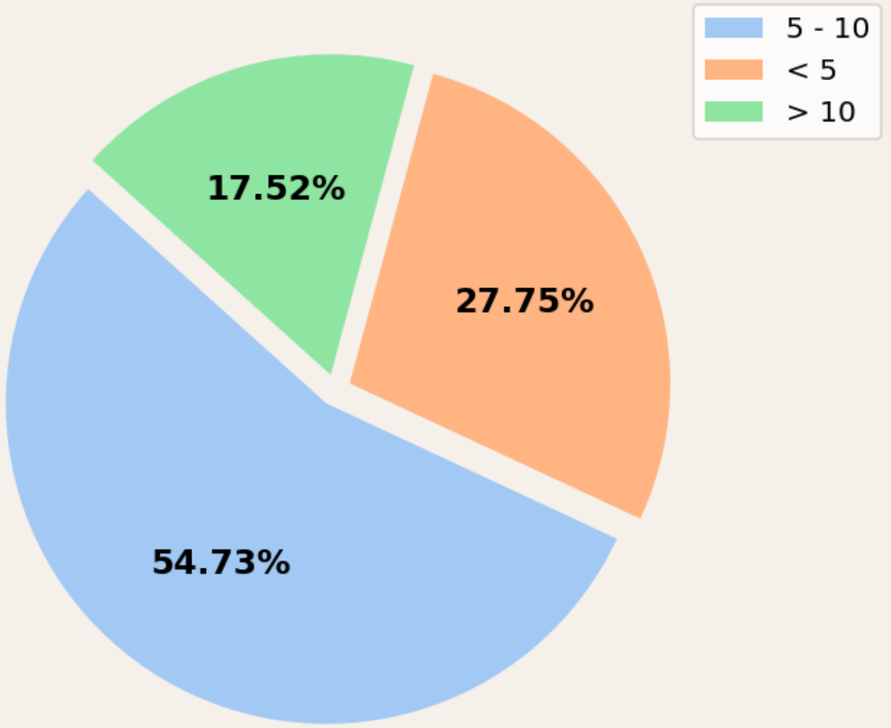


Analisis exploratorio

Primer hijo

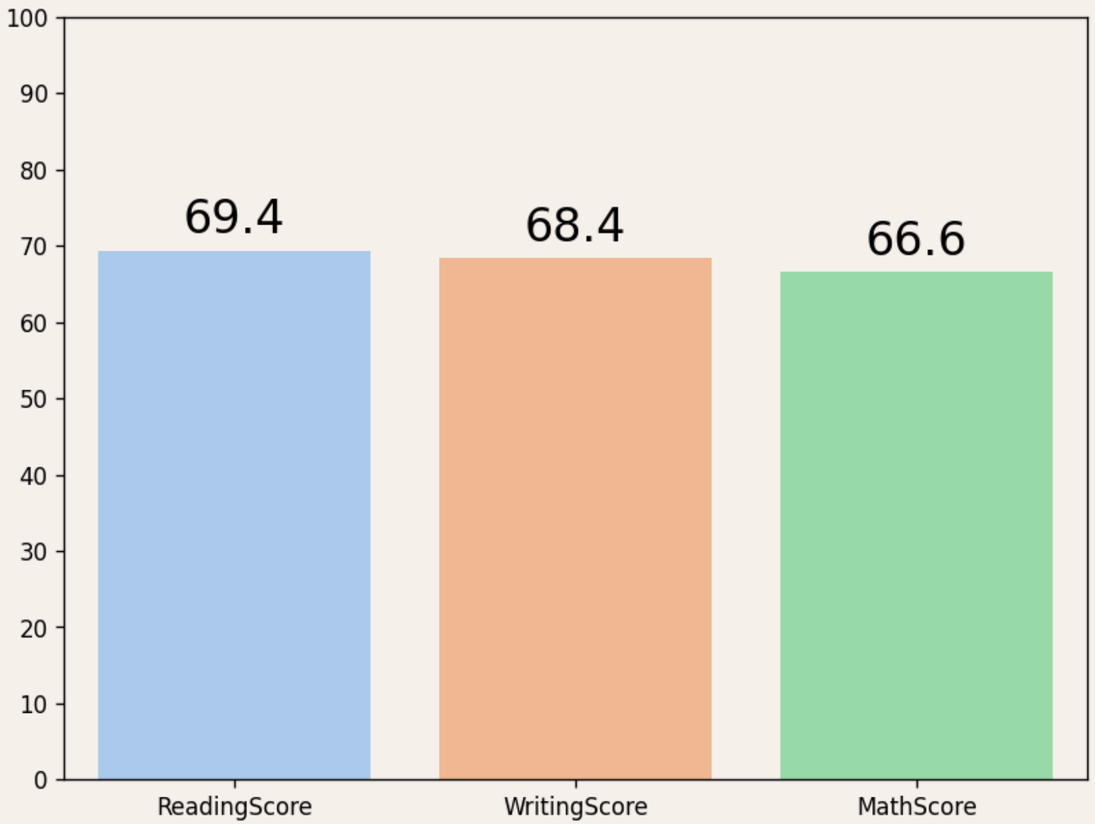


Horas semanales de estudio



Analisis exploratorio

Media de notas



¿Matemáticas con la media más baja?

Preparacion de los datos

Del analisis exploratorio decidimos:

- Remplazar los valores faltantes con SimpleImputer y utilizando la estrategia most frequent.

```
['female' 'group C' 'some college' 'standard' 'none' 'married'
 'sometimes' 'yes' 1.0 'school_bus' '5 - 10' 64 65 67]
```

- Valores categoricos: Codificacion por OneHotEncoding

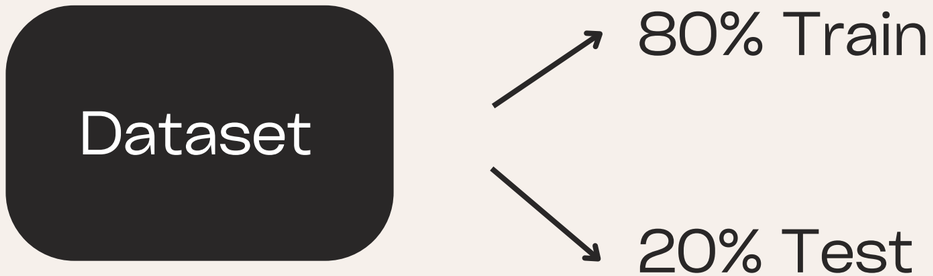
```
['Gender', 'EthnicGroup', 'ParentEduc', 'LunchType', 'TestPrep',
 'ParentMaritalStatus', 'PracticeSport', 'IsFirstChild',
 'TransportMeans', 'WklyStudyHours']
```

Preparacion de los datos

Del analisis exploratorio decidimos:

- Crear una unica columna:
  - Hallando el promedio entre las 3 notas, si es mayor o igual a 6 se considera aprobado y se coloca 1, si es menos se coloca un 0

- También se realizo una división de los datos en conjuntos de Train y Test de la siguiente forma



Modelado

Modelo a utilizar

Polynomial features

↓

Standar Scaler

↓

Logistic Regression

Utilizaremos un pipeline para realizar varios pasos de procesamiento y modelado

- Polynomial features: Para generar caracteristicas polinomiales
- Standar Scaler: Para realizar un escalado de características y asegurar una media de 0 y desviación de 1
- Logistic Regression: Implementa un algoritmo de regresión logística. max_iter de 2000

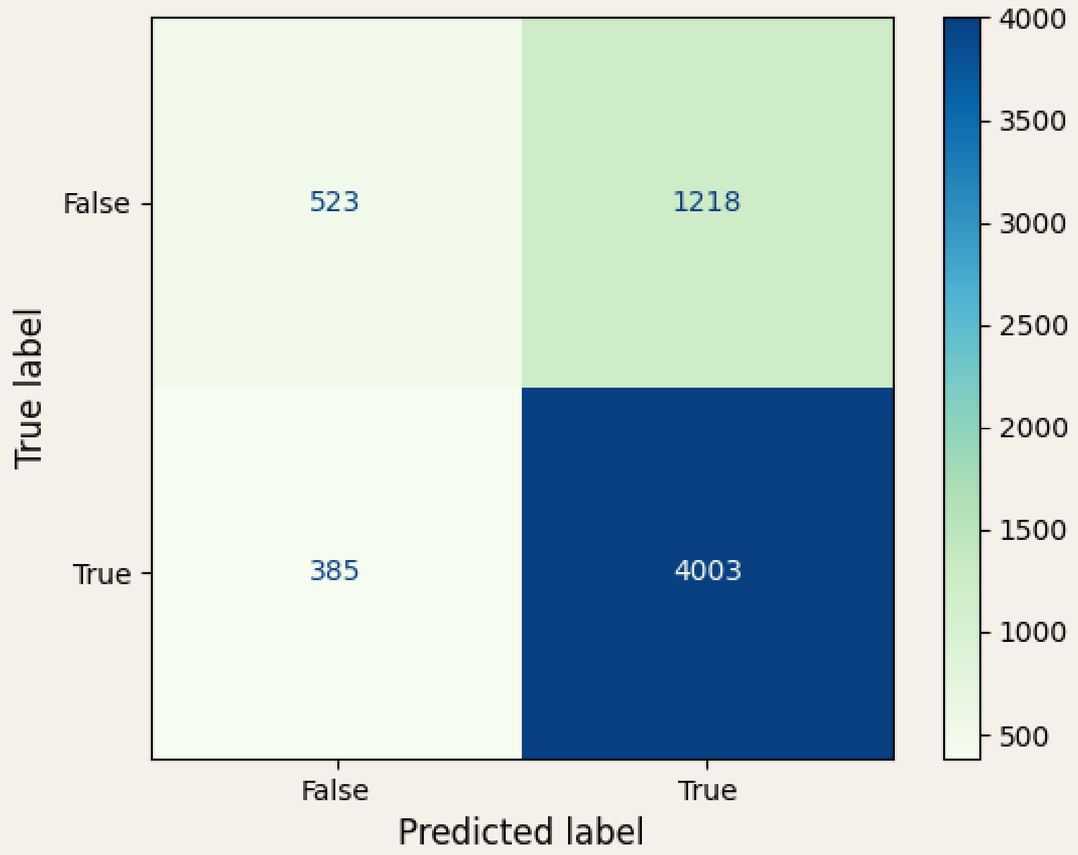
Determinacion de hiper-parametros

Se utilizo gridsearch para elegir los mejores hiper-parámetros

{'logi__C': 0.1 'poli__degree': 1}

Metricas

Matriz de confusion



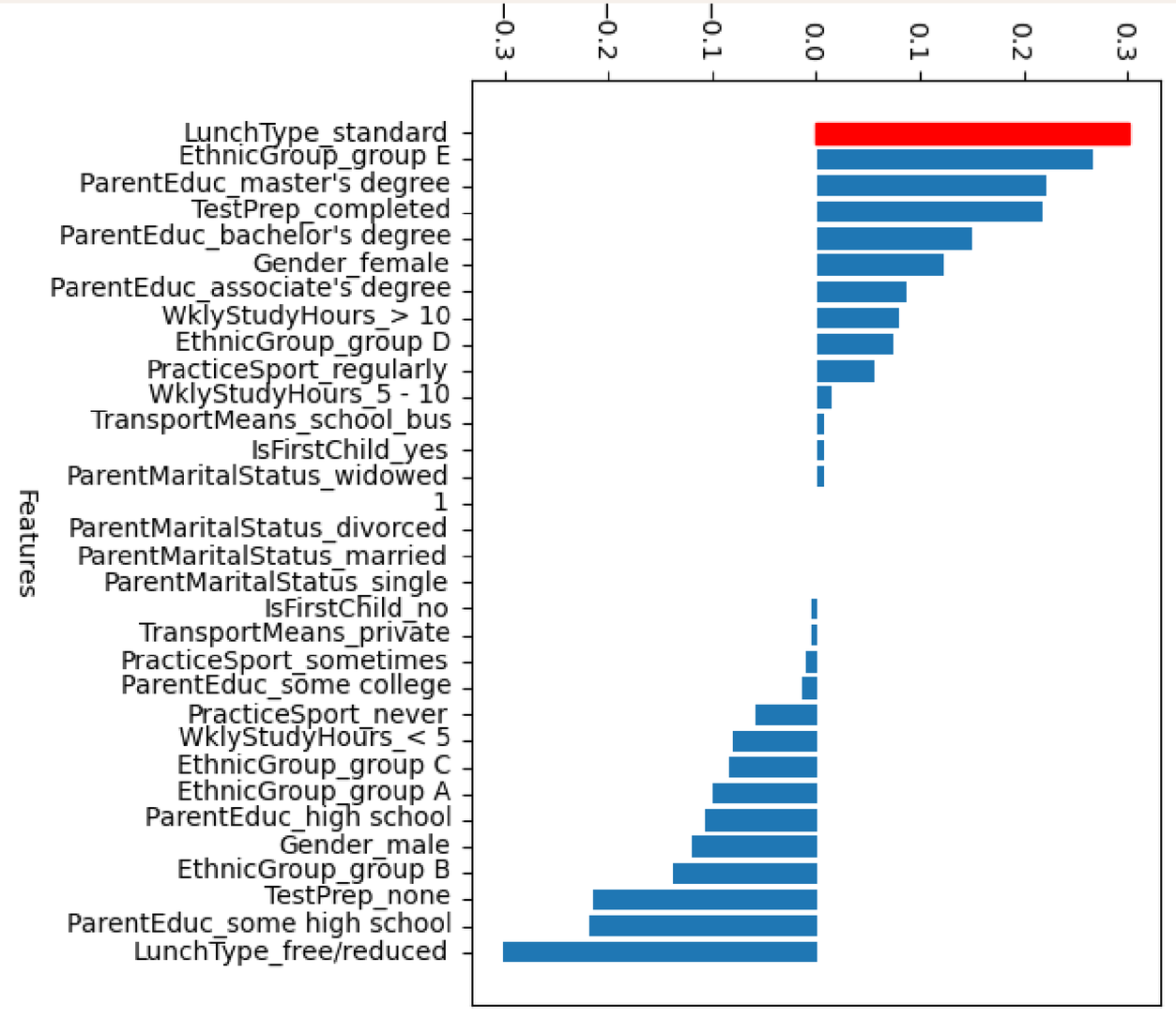
Accuracy: 0.738456518192201

Precicion: 0.766711357977399

F1-Score: 0.8331772296805078

Recall: 0.912260711030082

Metricas



Features más importantes

Conclusiones

El modelo tiene un buen desempeño y nos da información sobre features mas importantes

En cuanto a si los modelos responden la pregunta planteada, señalan algunas características influyentes: tipo de almuerzo, educación de los padres, preparación al test, horas semanales de estudio y si practica deporte

Algunas acciones a realizar para mejorar el rendimiento académico entre los alumnos es asegurar la buena alimentación, realizar deporte y fomentar hábitos de estudio

Muchas gracias