Ánalisis de las Calificaciones en Estados Unidos

Lenin Escalante
420003193

Mayo 2023

1 Introducción

En este proyecto se analizará el desempeño académico de 1000 estudiantes en EE.UU en 3 exámenes que miden sus habilidades en matemáticas, lectura y escritura tomando en cuenta el trasfondo de cada alumno, incluyendo factores tales como la escolaridad máxima de los padres o si el alumno recibió apoyo alimentario en la escuela, con el fin de identificar factores importantes que influyan de manera positiva en el resultado de los alumnos

2 Descripción breve de los datos y objetivo

Los datos a analizar constan de una tabla con 1000 registros obtenida de Kaggle, cada registro representa a un alumno, el objetivo de este proyecto será crear un modelo que clasifique el rendimiento de los alumnos como bueno o malo, para esto, se promediarán los resultados de las 3 pruebas para cada alumno, las calificaiones de las 3 pruebas van del 0 al 100.

El cuantil $\alpha = 0.65$ del promedio general de todos los alumnos es de 74.67, se redondeó este valor a 75 y para este proyecto, se define que un alumno tuvo un buen resultado si su promedio general fue mayor o igual a 75, y malo en caso contrario

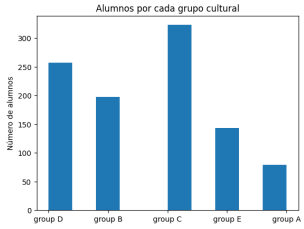
3 Análisis Exploratorio

Se importaron los datos de la tabla Exams a SQL usando pgAdmin4 para la realización del análisis exploratorio previo.

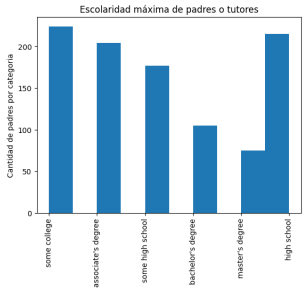
Las primeras observaciones son las siguientes: existen 1000 registros en total, cada registro representa a un alumno, además no existe identificador único para cada alumno. Ninguna de las columnas de la tabla cuenta con valores nulos. A continuación se enlista cada una de las columnas y una breve descripción de su contenido:

- Gender: Genero de cada estudiante (masculino o femenino)
- Ethnicity: Raza o grupo cultural del estudiante, existen 5 categorías
- Parent Level of Education: Escolaridad maxima de padres o tutores, existen 6 categorías
- Lunch: Indica si el alumno recibe o no apoyos a la alimentación
- Test preparation course: Indica si el alumno terminó o no un curso de preparación previo al examen (sí/no)
- Math score: puntaje obtenido en la evaluación de matemáticas, de 0 a 100
- Reading score: puntaje obtenido en la evaluación de lectura, de 0 a 100
- Writing score: puntaje obtenido en la evaluación de redacción, de 0 a 100

Respecto a la columna Gender, de los 1000 aplicantes a las 3 evaluaciones, 508 son hombres y 492 son mujeres. En cuanto a la categorización en razas; existen 5 grupos en los cuales se clasifica a cada alumno (A,B,C,D y E), la distribución de alumnos por grupos es la siguiente:

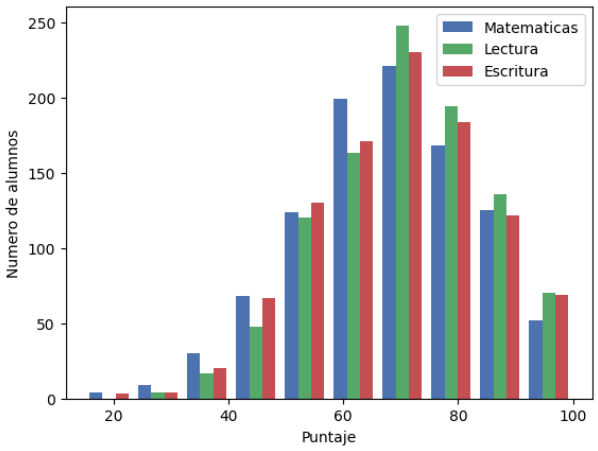


La escolaridad de los padres o tutores se divide en 6 categorías: *master's degree, some college, bachelor´s degree, high school* y *some highscool*, y la cantidad de alumnos en cada categoría se distribuye como puede verse en el siguiente histograma



Pasando al tema de apoyos alimentarios, 340 alumnos reciben un apoyo parcial o total en su alimentación, mientras que los 660 alumnos restantes no reciben ningún tipo de apoyo. Los datos acerca de la preparación de los estudiantes muestran que 656 no completaron el curso previo a las pruebas, los 334 restantes sí lo hicieron.

Las calificaciones de cada una de las 3 pruebas se distribuyen de la siguiente manera:



Como se puede observar, las calificaciones obtenidas en cada una de las 3 pruebas se distribuyen de manera similar, las medianas de las calificaciones de matemáticas, lectura y escritura son, respectivamente:

- Matemáticas: 68
- Lectura: 70.5
- Escritura: 70

Mientras que los promedios, siguiendo el orden establecido, son de 67.81, 70.38 y 69.14. Las calificaciones mínimas en cada examen fueron de: 15, 25 y 15; en los 3 exámenes la máxima calificación obtenida fue de 100.

4 Tabla ABT

Para la construcción de la tabla ABT, con la que se realizará un modelo de clasificación del rendimiento de cada estudiante, se tomaron las variables existentes del conjunto de datos original, haciendo las modificaciones pertinentes a las variables categóricas, y se define una variable adicional como el promedio general de un alumno en las 3 evaluaciones, es decir:

$$\text{Promedio general} = \frac{\text{Matemáticas} + \text{Lectura} + \text{Escritura}}{3}$$

Empleando la variable recien creada, se define la variable target de la siguiente manera

$$F(x) = \begin{cases} 1 & \text{Promedio general} \geq 75 \\ 0 & \text{Promedio general} < 75 \end{cases} \tag{1}$$

En resumen, la tabla ABT cuenta con todas las variables originales, y a continuación se listan los cambios hechos a algunas variables y la definición de las nuevas

- Gender: dividida en variable dummy [1]
- Ethnicity: dividida en variable dummy
- Parent level of education: dividida en variable dummy
- Lunch: dividida en variable dummy
- Test preparation course: dividida en variable dummy
- Promedio general: Promedia las calificaciones obtenidas en las 3 pruebas
- Variable objetivo: 1 si la el promedio general es mayor o igual a 75, 0 en caso contrario

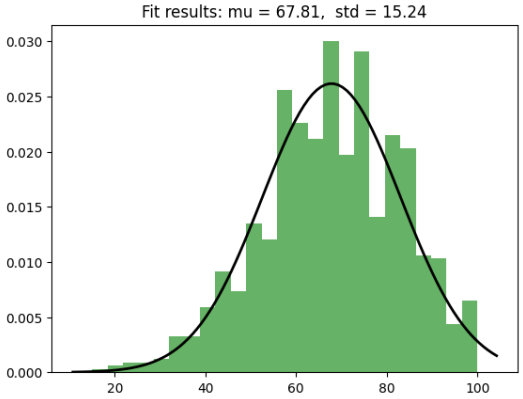
Las demás variables permanecen sin cambios

[1] Una variable dummy en el análisis de regresión es una variable binaria que se utiliza para representar categorías o grupos distintos. Toma el valor de 1 si una observación pertenece a un grupo en particular y 0 si no lo hace

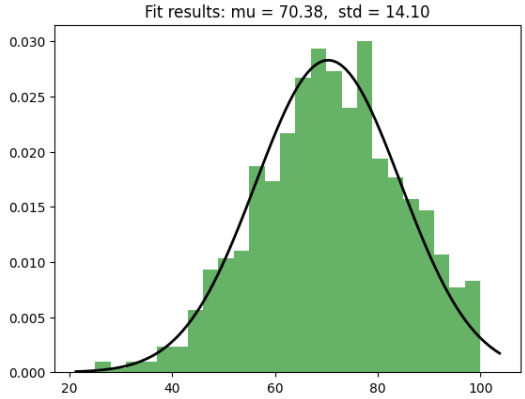
5 Análisis Estadístico

5.1 Distribución de variables numéricas

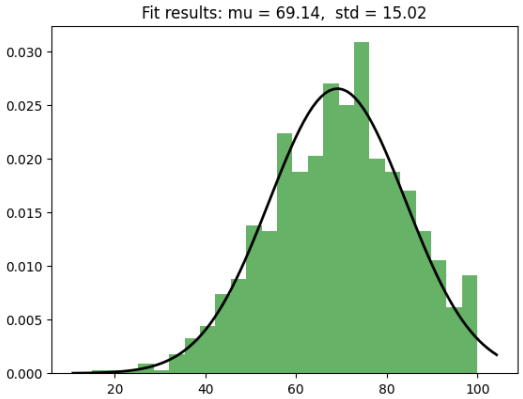
Del histograma hecho en el análisis exploratorio, se observó que las calificaciones de los 3 exámenes tienen distribuciones similares; se realizó la prueba de Kolmogorov-Smirnoff [2] a la distribucion de calificaciones de los 3 exámenes y también al promedio general, tomando un nivel de significancia $\alpha = 0.05$; se obtuvo que las variables tienen las siguientes distribuciones:



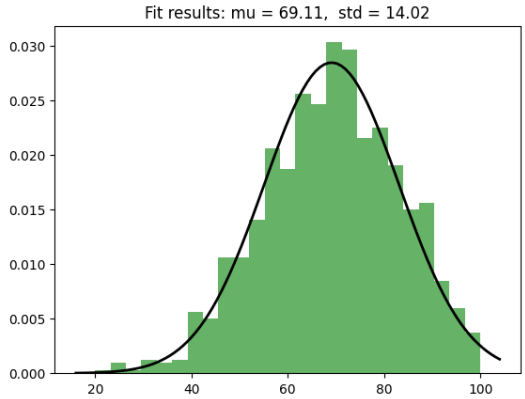
(a) Math Score, $N(67.81, 15.24^2)$



(b) Reading score, $N(70.38, 14.10^2)$



(c) Writing Score, $N(69.14, 15.02^2)$



(d) Average score, $N(69.11, 14.02^2)$

Por lo que, se puede asumir que las 4 distribuciones son normales, con parámetros bastante similares ya que los p-values de cada prueba fueron menores al nivel de significancia $\alpha = 0.05$ [3]

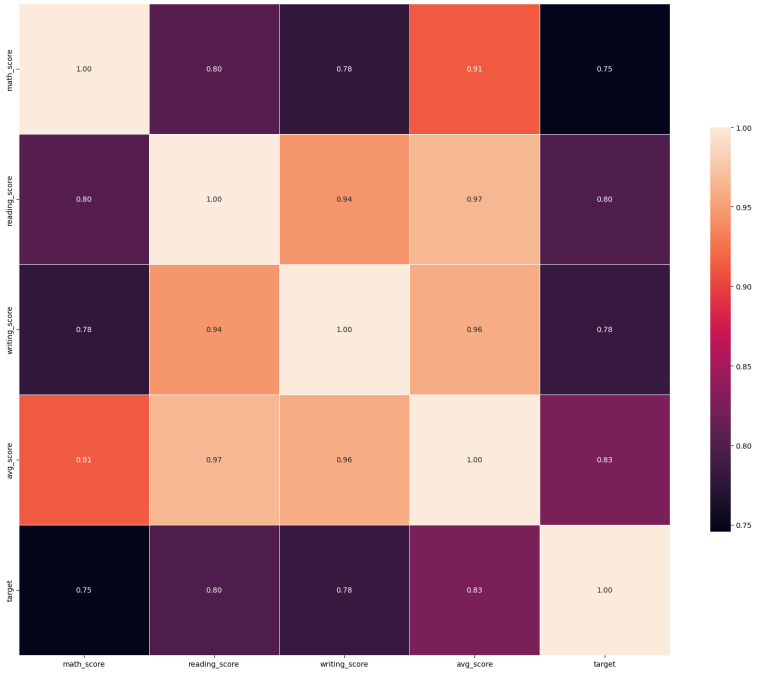
[2] La prueba de Kolmogorov-Smirnoff se usa para determinar que tan bien cierta distribución se ajusta a los datos, siendo la hipótesis nula que la muestra sigue una distribucion F propuesta, y la alternativa el caso contrario, en este caso la hipótesis nula es: Muestra $\sim N(\mu, \sigma^2)$

[3] La regla de decisión del p-value en pruebas de hipótesis establece que si el p-value asociado a una prueba estadística es menor que el nivel de significancia predefinido (generalmente $\alpha = 0.05$), se rechaza la hipótesis nula y se concluye que hay evidencia suficiente para respaldar la hipótesis alternativa.

5.2 Análisis de Correlación

5.2.1 Variables numéricas

Se utilizó el coeficiente de correlación de Spearman, El coeficiente de correlación de Spearman evalúa la relación entre dos variables ordinales. Va de -1 a 1, donde 1 es una correlación positiva perfecta, -1 es una correlación negativa perfecta y 0 indica ausencia de correlación. Esta medida se utilizó para analizar la correlación entre las variables: math score, reading score, writing score, average score y la variable target:



Como se aprecia del mapa de calor, las 4 variables están ampliamente correlacionadas entre sí, incluyendo la variable objetivo, por lo que sería una buena idea remover estas variables de nuestro modelo, para evitar que el modelo llegue a conclusiones incorrectas

5.2.2 Variables categóricas

Se realizo la prueba de la Chi Cuadrada para determinar la correlación entre variables categóricas, esta prueba no paramétrica se utiliza para determinar dependencia en variables categóricas, siendo la hipótesis nula que las variables son independientes. Se utilizó esta prueba entre las variables originales, comparándolas entre sí; y luego se realizó para determinar si existe una relación entre las variables originales y target.

Al nivel de significancia $\alpha = 0.05$, se obtuvieron los siguientes resultados:

- Correlación entre variables originales: Todas las variables categóricas originales son independientes entre sí; a excepción de lunch y test preparation
- Correlación entre variables originales y target: Las variables categóricas originales están bastante relacionadas con la variable target, con p-values mucho menores a $\alpha = 0.05$

5.2.3 Correlaciones Mixtas

A continuación se compararon las variables gender y lunch con las 4 variables numéricas que hemos definidas hasta ahora, para esto; se utilizó el coeficiente de correlación biserial puntual, el cuál puede ser usado para determinar la correlación entre variables categóricas nominales y variables continuas de tipo normal (como en este caso), los resultados, tomando un p-value de $\alpha = 0.05$ fueron los siguientes:

- Existe una correlación negativa entre el género femenino y el rendimiento en el examen de matemáticas, con un coeficiente de correlación de -0.19
- Existe una correlación positiva entre el genero femenino y el rendimiento en las pruebas de lectura y en redacción, con coeficientes de correlación de 0.21 y 0.28
- El género femenino tiene pequeña una correlación positiva con el promedio general, con un coeficiente de 0.1
- Recibir apoyos alimenticios está negativamente correlacionado con los resultados en los 3 exámenes y con el promedio general, con coeficientes de correlación que van de desde -0.23 a -0.37
- La preparación previa está positivamente relacionada con los resultados en las 3 evaluaciones y con el promedio general

5.2.4 Conclusiones Generales

Del análisis estadístico se resaltan las siguientes observaciones:

- Los resultados de los 3 exámenes, y el promedio general, tienen una distribución normal
- Existe una correlación entre si el alumno recibe apoyos alimentarios y concluyo la prueba de preparación
- Las mujeres tienden a tener peores resultados en matemáticas en comparación con los hombres, pero en las demás pruebas tienen un mejor desempeño, al igual que en el promedio general
- Recibir cualquier tipo de apoyos alimentarios está correlacionado a un rendimiento menor en las pruebas
- Completar la preparación previa al examen está correlacionado con mejores resultados en las pruebas
- El grupo étnico E está altamente correlacionado con mejores resultados, al igual que la escolaridad máxima de los padres [4]
- Las variables de puntaje en las 3 evaluaciones, y el promedio general están muy relacionadas con la variable objetivo, por lo que serán removidas de este modelo

Se espera que los modelos revelen tendencias similares entre los datos y la variable objetivo

6 Modelación

Se utilizaron algoritmos de clasificación para pronosticar, según las variables creadas en la tabla ABT, si un estudiante obtuvo un buen promedio general en los 3 exámenes; usando la variable target. Dicha variable se definió como 1 si su promedio era mayor o igual a 75, y 0 en caso contrario.

6.0.1 Resumen de modelos utilizados

Este es un problema de clasificación, se usaron los siguientes algoritmos de aprendizaje supervisado para problemas de clasificación

- Naive Bayes: algoritmo de clasificación que se basa en el teorema de Bayes y asume una independencia entre las características o variables predictoras, Se utiliza para asignar una clase o categoría a una nueva instancia según la probabilidad condicional de esa clase dada la instancia.
- Modelo Lineal Generalizado: El modelo lineal generalizado (GLM) para clasificación es una técnica estadística que se utiliza para la clasificación de datos. Es una extensión del modelo lineal tradicional que permite trabajar con variables de respuesta categóricas o binarias. En lugar de ajustar una línea recta, el GLM utiliza una función de enlace para modelar la relación entre las variables predictoras y la probabilidad de pertenecer a una clase específica.

[4] La matriz de correlación completa está disponible en el notebook de python

- Regresión Logística: La regresión logística es un algoritmo de clasificación utilizado para predecir la probabilidad de pertenencia a una clase binaria. Utiliza una función logística para modelar la relación entre las variables predictoras y la probabilidad de pertenecer a la clase positiva.

- Fast Large Margin: Fast Large Margin (FLM) es un enfoque de clasificación que se basa en el concepto de márgenes amplios. El objetivo es encontrar un hiperplano de separación entre clases que maximice la distancia entre los puntos de datos y el hiperplano. FLM es útil cuando se necesita un clasificador de alta precisión y robusto ante datos no lineales o ruidosos.

- Red Neuronal: Una red neuronal para clasificación es un modelo de aprendizaje automático inspirado en el funcionamiento del cerebro humano. En el contexto de la clasificación, una red neuronal aprende a reconocer patrones y características en los datos de entrada para asignarlos a diferentes clases.

- Arbol de Decisión: El árbol de decisión para clasificación es un modelo que utiliza una estructura de árbol para tomar decisiones de clasificación basadas en características de los datos. Cada nodo representa una pregunta y cada rama es una respuesta. Las hojas del árbol son las clases a las que se asignan los datos. Es un modelo interpretativo y eficaz para clasificar datos numéricos y categóricos.

- Bosque Aleatorio: El bosque aleatorio para clasificación es un modelo que combina múltiples árboles de decisión para mejorar la precisión de la clasificación. Cada árbol se entrena con diferentes subconjuntos de datos y características. Luego, las predicciones de todos los árboles se combinan para obtener una predicción final mediante votación o promediado. El bosque aleatorio es efectivo para clasificar datos complejos, reduce el sobreajuste y es capaz de manejar grandes conjuntos de datos.

- Gradient Boosted Trees: El bosque aleatorio con Gradient Boosting para clasificación es un modelo que combina múltiples árboles de decisión utilizando el enfoque de Gradient Boosting. En lugar de entrenar los árboles de forma independiente, el Gradient Boosting ajusta cada árbol en función de los errores del modelo anterior, priorizando las instancias clasificadas incorrectamente.

- Support Vector Machine (SVM): Algoritmo de aprendizaje automático que se utiliza para separar datos en diferentes categorías. Busca encontrar un hiperplano que maximice el margen entre las muestras de diferentes clases. Es efectivo en conjuntos de datos pequeños a medianos y ha demostrado ser útil en problemas de clasificación binaria y multiclase.

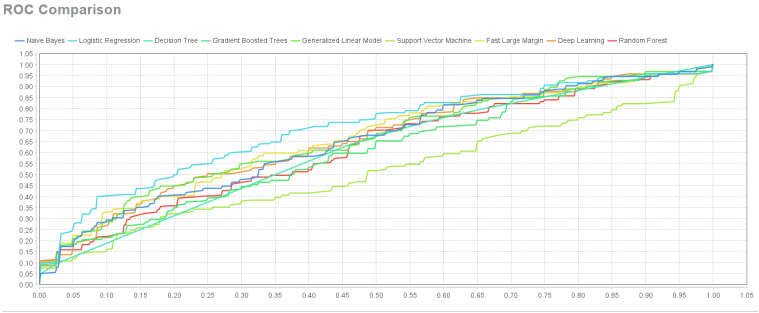
Los algoritmos de árbol de decisión y SVM fueron implementados en Python, mientras que los demás fueron implementados usando la herramienta Rapidminer.

6.0.2 Rendimiento de cada modelo

Abajo se muestra la curva ROC de cada modelo y su correspondiente area bajo la curva ROC, la curva ROC Representa la relación entre la tasa de verdaderos positivos y la tasa de falsos positivos. Una curva ROC ideal se acerca al vértice superior izquierdo, lo que indica una alta tasa de verdaderos positivos y una baja tasa de falsos positivos.

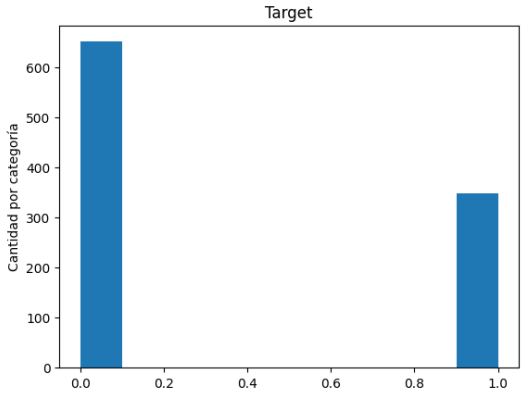
El área bajo la curva ROC es una métrica común utilizada para cuantificar el rendimiento del modelo, donde un valor de 1 representa una clasificación perfecta y 0.5 indica una clasificación aleatoria.

Modelo	AUC
Naive Bayes	0.65
Generalized Linear Model	0.66
Logistic Regression	0.71
Fast Large Margin	0.66
Deep Learning	0.66
Decision Tree	0.61
Random Forest	0.62
Gradient Boosted Trees	0.61
Support Vector Machine	0.51



El mejor algoritmo tomando como métrica el área bajo la curva ROC es la regresión logistica, con un puntaje de 0.71 en esta métrica.

SE debe tomar en cuenta que existe un ligero desbalance en la variable target, como puede verse en la siguiente grafica:



Tomando esto en cuenta, sería conveniente usar el Coeficiente de Correlación de Matthew, esta métrica es preferible en este caso ya que toma en cuenta los verdaderos positivos, los verdaderos negativos, los falsos negativos y los falsos positivos. Un coeficiente de correlación de Matthew alto indica que nuestro modelo predice bien ambas categorías. Abajo se muestra la matriz de confusión de cada uno de los modelos, con su correspondiente coeficiente de Matthew:

Modelos	Matriz de confusión			correlación de Mathew
Naive bayes		true range1	true range2	0.202842627
	pred. range1	180	91	
	pred. range2	3	11	
Modelo lineal generalizado		true range1	true range2	0.240610992
	pred. range1	179	87	
	pred. range2	4	15	
Regresion Logística		true range1	true range2	0.268814235
	pred. range1	176	85	
	pred. range2	5	19	
FLM		true range1	true range2	0.260101177
	pred. range1	177	83	
	pred. range2	6	19	
Red Neuronal		true range1	true range2	0.115228698
	pred. range1	186	97	
	pred. range2	0	2	
Arbol de Decision		true range1	true range2	0.094105802
	pred. range1	195	62	
	pred. range2	96	47	
Bosque Aleatorio		true range1	true range2	0.081335418
	pred. range1	186	98	
	pred. range2	0	1	
Gradient Boosted Trees		true range1	true range2	0.227419754
	pred. range1	179	88	
	pred. range2	4	14	
SVM		true range1	true range2	0.293118824
	pred. range1	227	30	
	pred. range2	91	52	

De la tabla, se puede apreciar que el modelo con mejor desempeño con esta métrica es el de regresión logística

7 Validación

Tomando en cuenta los resultados del análisis de correlación, empleando la tabla ABT seleccionaremos un subconjunto de esta que cumpla las siguientes características:

- Los alumnos deben de haber completado el curso de preparación para el examen
- Los padres deben de tener una escolaridad máxima mayor a la de secundaria
- No deben recibir apoyos alimentarios
- Los alumnos deben pertenecer a la raza E

Se realizó la matriz de confusión de este subconjunto de datos empleando el modelo ya entrenado de regresión logística, ya que es el que obtuvo una mejor metrica de Matthew

	True Buen Resultado	True Mal Resultado
Pronosticado Buen Resultado	10	4
Pronosticado Mal Resultado	2	33

Como se puede observar, en este subconjunto de datos, la matriz de confusión es muchísimo mejor que tomando toda la tabla ABT, y el coeficiente de correlación de Matthew en este caso es de: 0.69, este coeficiente tambíen es muchísimo más alto en este caso.

Esto indica que factores importantes para determinar el desempeño de los alumnos son los considerados en el subconjunto de la tabla ABT creada.

Para corroborar estos resultados, se utilizó la herramienta Rapidminer, y se obtuvo que las siguientes variables son más importantes a la hora de determinar el buen desempeño de los alumnos:

También se determinó el peso que cada variable tiene en la variable objetivo, en la tabla de abajo se muestra cada variable con su correlación ordenada de mayor a menor. Esta tabla corrobora las observaciones hechas en esta sección (ver pagina siguiente).

Atributo	Correlación Con Target
lunch_standard	0.229
lunch_free/reduced	0.229
ethnicity_group E	0.198
test_preparation_completed	0.154
test_preparation_none	0.154
ethnicity_group B	0.090
parent_level_associate's degree	0.087
parent_level_some high school	0.081
ethnicity_group C	0.071
ethnicity_group A	0.067
gender_female	0.064
gender_male	0.064
parent_level_high school	0.056
parent_level_master's degree	0.046
ethnicity_group D	0.040
parent_level_bachelor's degree	0.023
parent_level_some college	0.001

8 Conclusiones

Aunque el modelo de regresión logística no tiene una metrica de Matthew alta, gracias al análisis estadístico realizado previo a la modelación se pudieron determinar algunos factores que podrían determinar el desempeño de los alumnos, y tras aplicar el filtro a la tabla ABT, pudimos verificar que el desempeño del modelo mejoró bastante. Esto arroja las siguientes conclusiones:

- La escolaridad máxima de los padres es un factor importante que determina el desempeño de sus hijos en la escuela, a menor escolaridad, peores resultados. La escolaridad minima para mejorar las posibilidades de obtener resultados favorables es tener la secundaria terminada
- Existe un grupo cultural que tiene un desempeño mucho mejor que los demás en todas las evaluaciones.
- El hecho de que el no recibir apoyos alimentarios esté relacionado a no completar la prueba previa al examen puede indicar que estos apoyos no son suficientes para garantizar que el alumno tenga las condiciones para estudiar
- Las mujeres tienen mejor rendimiento promedio que los hombres, aunque esta diferencia es mínima, el hecho de que los hombres tengan un mejor rendimiento en matemáticas pero peor rendimiento en las demás pruebas podría implicar que el enfoque del sistema educativo crea disparidades en el rendimiento de cada género.

El análisis arrojó resultados interesantes que podrían ser utilizados para mejorar la calidad de la educación en tres áreas que son básicas para el desarrollo adecuado de los alumnos, aún tomando en cuenta las fallas de los modelos, se obtuvieron conclusiones bastante importantes para el problema a resolver.