Examen Mineduc, Graduandos 2017

Sergio Vinicio García Monterroso

Universidad Francisco Marroquín

24 de abril, 2019

Introducción

Existen registros públicos digitales que datan desde 1998 sobre los exámenes que se les hacen a los alumnos que cursan su último año del colegio por parte del Mineduc. Aparte de su rendimiento en el examen de lectura y matemática, se recolectan datos por medio de una encuesta a todos los participantes. Esta encuesta se efectúa antes de empezar el examen y son datos generales de los estudiantes que van desde el tipo de vivienda hasta como pasan su tiempo libre.

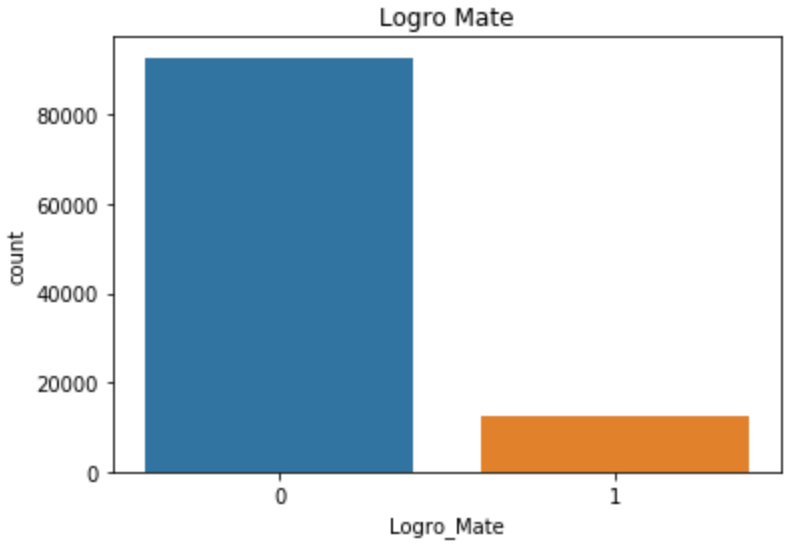
El objetivo del presente modelo de clasificación no es solamente lograr estimar de manera adecuada el rendimiento que obtendrán los alumnos basado en el fondo que hay detrás de cada uno de ellos, sino que tiene como objetivo compartido lograr identificar las variables que mas afectan el rendimiento en el examen. La educación es algo fundamental en el desarrollo integral de un país, si se logran identificar factores que puedan mejorar el mismo, entonces se podría trabajar por un mejor sistema educativo o bien mejorar las condiciones de vida que promueven a un buen rendimiento por parte de los alumnos.

Los mayores retos parar lograr un modelo que cumpla con estos objetivos son:

- Limpiar los datos que otorga el MINEDUC.
- Elegir el modelo que mejor se adapte a los datos.

Para este modelo de clasificación se intentará clasificar de manera adecuada el logro de cada alumno en el examen de matemáticas. Se decidió hacerlo con el examen de matemáticas y no el de lectura ya que existe una correlación fuerte entre el primero y el segundo. Se demostró que quienes ganan matemáticas ganan el de lectura, pero no al revés.

Para entender mejor la situación del país basta con ver la siguiente gráfica.



La cantidad de alumnos que logra ganar el examen de matemáticas es mínima comparando con los que pierden. Este tipo de datos refleja la realidad nacional, y resalta la importancia urgente de mejorar las condiciones educativas de los niños guatemaltecos.

Datos

Para este análisis se tomaron los datos del año 2017. Los mismos cuentan con 199 variables y 158,962 observaciones que provienen de alumnos de 4,124 establecimientos distintos.

Limpieza

El proceso de limpieza de estos datos fue muy extenso, ya que se encontraban demasiados Nan, datos irrelevantes para el modelo, y redundancia en las variables. Por ejemplo, para determinar de que departamento era un alumno. En vez de existir una variable "dummy" para cada uno de los departamentos, existían 2 variables una que le otorgaba un número a cada departamento y otra en donde ponían el nombre de cada departamento. En este tipo de casos, de la variable de que tenía el nombre de cada departamento se sacaban 22 variables "dummies" para cada departamento. Y después, se eliminaban las 2 variables anteriormente mencionadas. También existían variables que ponían como observación "9 = No sabe". Este tipo de valores se omitieron, ya que no contribuyen en nada al modelo.

Por otra parte, no se encontraron valores atípicos. Esto es algo entendible ya que se trata de una encuesta y no deberían existir este tipo de valores, a no ser que haya errores a la hora de ingresar los datos.

Después de todo el proceso de limpieza, se quedó finalmente con un dataset que cuenta con 185 variables y 105,322 observaciones.

Métodos

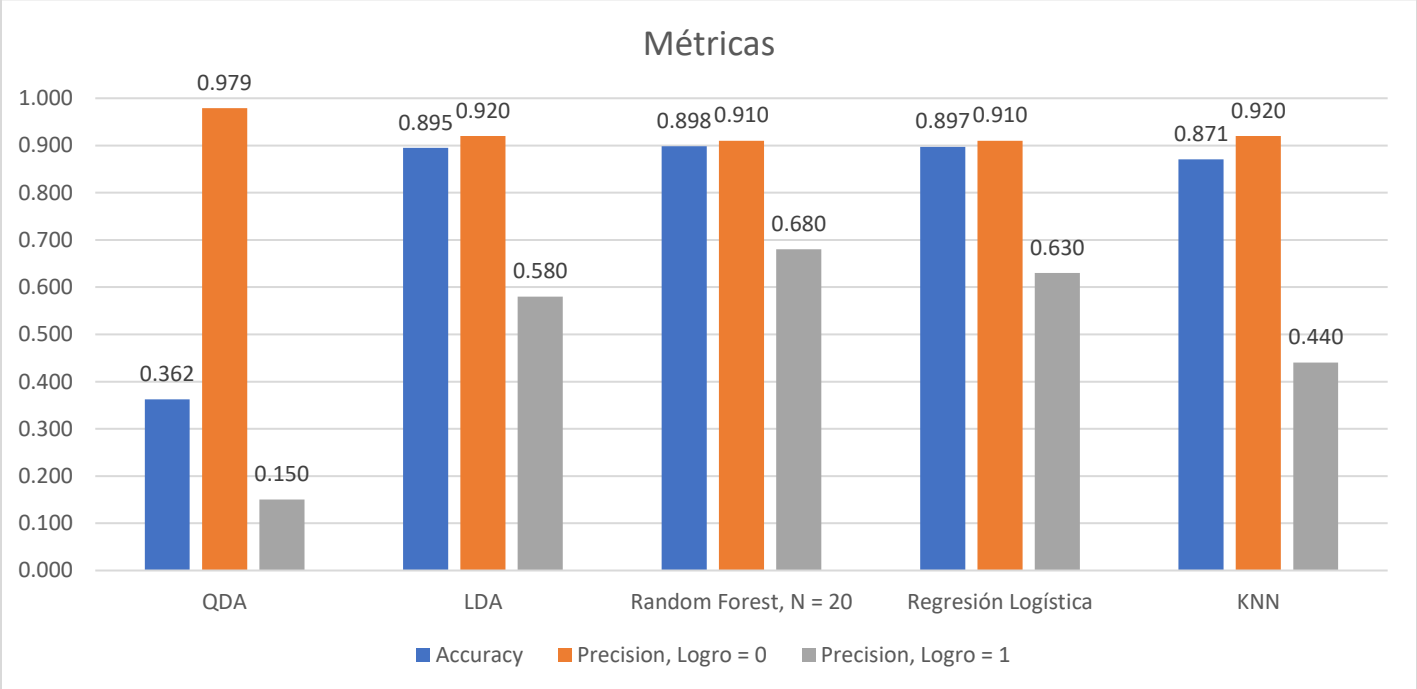
Con el fin de encontrar el mejor modelo de clasificación se corrieron 5 modelos distintos. A partir de esos 5, se agarró el mejor y se buscó mejorarlo y también reducir la cantidad de variables para identificar a las más relevantes.

Los 5 algoritmos utilizados fueron los siguientes:

1. QDA, Análisis discriminante cuadrático
2. LDA, Análisis discriminante lineal
3. Random Forest
4. Regresión Logística
5. KNN, K vecinos más cercanos

Resultados

Como primer punto para elegir al mejor modelo de los 5 se utilizaron 2 métricas principales "Accuracy" y "Precision". "Accuracy" es el porcentaje de clasificaciones correctas del total. Mientras que "precision" es la capacidad de un clasificador para no etiquetar una instancia como positiva que en realidad es negativa. Para cada clase se define como la proporción de verdaderos positivos con respecto a la suma de verdaderos y falsos positivos. Matemáticamente, precision se calcula de la siguiente manera "Precision = TP/(TP + FP)". Siendo TP los verdaderos positivos y FP los falsos positivos.



Estas métricas fueron obtenidas evaluando el modelo con el conjunto de prueba. De las 185 variables y 105,322 observaciones, se sacaron 2 conjuntos uno de entrenamiento y otro de prueba. Se utilizó una proporción de 75% para entrenamiento y 25% para prueba, quedando el conjunto de entrenamiento con 78,991 observaciones y el de prueba con 26,331.

Por el enfoque de esta investigación la métrica mas importante a evaluar es la de "Precision", Logro = 1, la cual representa la precisión que tuvo el modelo para predecir correctamente a los alumnos que ganarían el examen matemático. Las demás métricas sirven como complemento a esta.

Con estos resultados, se puede notar que claramente "Random Forest" supera a los demás modelos en dicha métrica. Además de ésto, se desempeña muy bien en las demás. Logrando un "accuracy" total del 0.898 y un "precision, Logro = 0" del 0.910. Esto quiere decir que del total de alumnos examinados lograr predecir con un acierto del 89.8% como les irá. Además, con 68.8% de acierto lograr clasificar correctamente a quienes ganarán y con un 91.0% a quienes perderán.

Modelo Final

Basado en los resultados anteriores, se intentó mejorar el modelo de "Random Forest" agregando más árboles, y se obtuvo una ligera mejora. Logrando un "accuracy" de 90.04%, un "precision" del 91% para quienes perderán, y un 71% para quienes ganarán. De dicho modelo se sacaron las variables mas influyentes, las cuales fueron:

- Periodos de matemática a la semana.
- Grado de educación que alcanzó el padre.
- Grado de educación que alcanzó la madre.
- Duración del periodo de matemáticas.
- Periodos en los que utiliza computadora.
- Edad.
- Duración del periodo de lectura.
- Periodos de lectura a la semana.
- Libros completos que ha leído por interés personal.
- Días que lee a la semana algún periódico.
- Horas diarias de uso de computadora en casa
- Horas diarias de uso de computadora en establecimiento

Seguido de esto, se corrió nuevamente el modelo "Random Forest" pero ahora solo con estas variables y el resultado fue el esperado. Se contó con un

“accuracy” del 89.32%, y aciertos del 91% y 60% para los que pierden y ganan su examen respectivamente.

Conclusión

Después de todo el trabajo e investigación se puede afirmar que los objetivos fueron cumplidos. Se cuenta con un modelo funcional que logra clasificar a buena parte de los examinados por parte del MINEDUC. El modelo tiene actualmente un 90% de acierto aproximadamente.

Aparte de eso queda documentado las variables mas influyentes en el desempeño de los alumnos en su examen. Se espera que esta investigación sirva como referencia para establecer un mejor sistema educativo y para que el gobierno o interesados en mejorar el sistema educativo nacional, puedan tener una base en donde deben invertir para mejorar el mismo.

Como posibles mejoras para este proyecto se propone explorar otros algoritmos que logren superar los resultados obtenidos, como redes neuronales. Aparte de eso, se considera que la base de datos se podría mejorar, ya que puede haber un posible sesgo por las preguntas efectuadas, entonces se propone implementar nuevas preguntas a la encuesta. Y finalmente, hacer una unificación de todas las bases de datos, por motivos de capacidad computacional en esta investigación se usaron datos solo del año 2017, se propone hacer este tipo de investigación unificando datos de todos los años disponibles.

Apéndice

- Datos
  - http://www.mineduc.gob.gt/digeduca/
- Métricas
  - https://muthu.co/understanding-the-classification-report-in-sklearn/
- Modelo
  - https://colab.research.google.com/drive/1bRlA3EwaMLRaoWJPftcGkMfLkJoDuXkP
- Limpieza de datos
  - https://github.com/SergioVinicio/Examenes-MINEDUC/blob/master/Vitales.R