Resultados Saber 11

Analítica Computacional para la Toma de Decisiones
Universidad de los Andes, Bogotá, Colombia

Santiago González Montealegre	s.gonzalez35	202012274
Juliana Carolina Cárdenas Barragán	jc.cardenasb1	202011683

Fecha de presentación: Junio 1 de 2023

Tabla de contenido

1 Introducción

Con base en la encuesta realizada durante el proceso de inscripción de los estudiantes en la prueba Saber 11 del ICFES, se ha emprendido la tarea de desarrollar un modelo de predicción de los resultados, centrándose en las condiciones socioeconómicas de los estudiantes. En el contexto colombiano, el Instituto Colombiano para la Evaluación de la Educación ha llevado a cabo pruebas de desempeño académico a nivel nacional, que permiten medir las capacidades de los estudiantes en diversas competencias académicas, con el objetivo de determinar su nivel educativo. Para este estudio, se ha utilizado una base de datos descargada del portal web de Datos Abiertos del Gobierno de Colombia. En este reporte se explica el análisis exploratorio realizado con el propósito de seleccionar las variables relevantes que puedan explicar el puntaje obtenido en la prueba y la creación, entrenamiento, testeo y evaluación de un modelo predictivo basado en redes bayesianas para predecir el resultado que obtendría un estudiante a partir de los valores ingresados.

2 Objetivo

El objetivo principal de este proyecto es crear una herramienta de análisis de datos destinada a los ciudadanos interesados en examinar el puntaje promedio de la prueba Saber 11 por departamento, estrato y ubicación del colegio (urbano o rural), y predecir el puntaje que un estudiante podría obtener en la prueba con base en información proporcionada.

3 Extracción de Datos

La tabla con los datos posee más de 50 atributos, por lo que resulta necesario segmentar cuales de los atributos son los necesarios para realizar nuestro modelo predictivo. Por lo que

primero se decide descargar los datos únicamente con las variables que se identificaron como pertenecientes a la red bayesiana. Primero se descargaron los datos crudos, posteriormente se subieron a un bucket en Amazon S3, luego se utilizo Amazon Glue para mapear los atributos y los datos a una tabla que pueda ser consultada por medio de Amazon Athena.

La primera consulta fue:

```
SELECT cole_area_ubicacion, cole_bilingue, cole_calendario, cole_naturaleza, cole_genero, cole_jornada, cole_cod_depto_ubicacion, cole_cod_mcpio_ubicacion, estu_cod_depto_presentacion, estu_cod_mcpio_presentacion, estu_cod_reside_depto, estu_cod_reside_mcpio, estu_genero, fami_estratovivienda, fami_educacionmadre, fami_educacionpadre, fami_personashogar, fami_tienecomputador, fami_tieneinternet, punt_global
FROM datos_crudos;
```

Con esta query ya hemos seleccionado las variables de interés, sin embargo, al hacer el Exploratory Data Analysis (EDA) nos encontramos con bastantes datos nulos. Por lo que es necesario otra query. A continuación, la query que filtra:

```
SELECT cole_area_ubicacion, cole_bilingue, cole_calendario, cole_naturaleza, cole_genero, cole_jornada, cole_cod_depto_ubicacion, cole_cod_mcpio_ubicacion, estu_cod_depto_presentacion, estu_cod_mcpio_presentacion, estu_cod_reside_depto, estu_cod_reside_mcpio, estu_genero, fami_estratovivienda, fami_educacionmadre, fami_educacionpadre, fami_personashogar, fami_tienecomputador, fami_tieneinternet, punt_global
FROM datos_crudos
WHERE COLE_AREA_UBICACION IS NOT NULL
AND COLE_AREA_UBICACION != '' AND COLE_BILINGUE IS NOT NULL AND COLE_BILINGUE != '' AND COLE_CALENDARIO IS NOT NULL AND COLE_CALENDARIO != '' AND COLE_NATURALEZA IS NOT NULL AND COLE_NATURALEZA != '' AND COLE_GENERO IS NOT NULL AND COLE_GENERO != '' AND COLE_JORNADA IS NOT NULL AND COLE_JORNADA != '' AND COLE_COD_DEPTO_UBICACION IS NOT NULL AND COLE_COD_MCPIO_UBICACION IS NOT NULL AND ESTU_COD_DEPTO_PRESENTACION IS NOT NULL AND ESTU_COD_MCPIO_PRESENTACION IS NOT NULL AND ESTU_COD_RESIDE_DEPTO IS NOT NULL AND ESTU_COD_RESIDE_MCPIO IS NOT NULL AND ESTU_GENERO IS NOT NULL AND ESTU_GENERO != '' AND FAMI_ESTRATOVIVIENDA IS NOT NULL AND FAMI_ESTRATOVIVIENDA != '' AND FAMI_EDUCACIONMADRE IS NOT NULL AND FAMI_EDUCACIONMADRE != '' AND FAMI_EDUCACIONPADRE IS NOT NULL AND FAMI_EDUCACIONPADRE != '' AND FAMI_PERSONASHOGAR IS NOT NULL AND FAMI_PERSONASHOGAR != '' AND FAMI_TIENECOMPUTADOR IS NOT NULL AND FAMI_TIENECOMPUTADOR != '' AND FAMI_TIENEINTERNET IS NOT NULL AND FAMI_TIENEINTERNET != '' AND PUNT_GLOBAL IS NOT NULL;
```

El resultado de esta query se guarda en una nueva tabla, llamada Transformed2. La mayoría de las variables son categóricas, por lo que en el EDA empezamos a ver los valores únicos que puede tomar cada atributo y nos encontramos con que hay viviendas que no cuentan con estrato, y hay padres y madres con valores de educación de no aplica o no sabe. Por lo que también filtramos estos valores en una nueva query sobre la tabla nueva:

```
SELECT *
FROM TRANSFORMED2
WHERE FAMI_ESTRATOVIVIENDA NOT IN ('Sin Estrato')
AND FAMI_EDUCACIONMADRE NOT IN ('No Aplica', 'No sabe')
AND FAMI_EDUCACIONPADRE NOT IN ('No Aplica', 'No sabe');
```

Los resultados de esta query se guardan en una nueva tabla, esta tabla será la que se descargará para poder realizar el EDA y la construcción del modelo predictivo

4 Análisis Exploratorio de Datos

Con la extracción anteriormente realizada se logra reducir la complejidad original de los datos con más 50 atributos y más de 7 millones de registros a unos datos más manejables con solo 20 atributos y 3 millones de registros.

Dentro del EDA se realizaron los siguientes gráficos para comparar la variable de interés puntaje global con las demás variables:

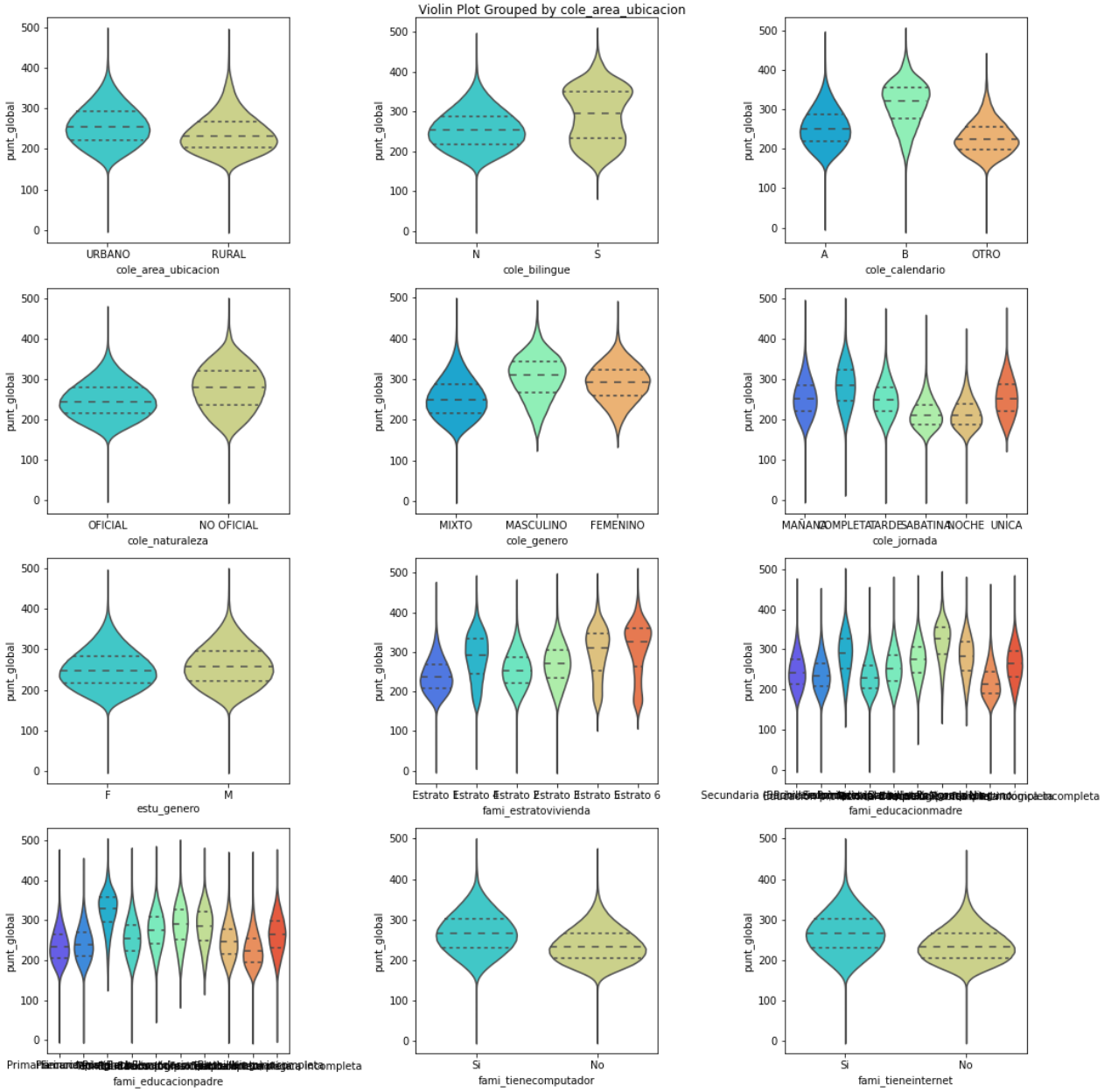

Figura 1. Puntaje vs Categoricas

También, se realizó un análisis para ver posibles gráficas de interés para los estudiantes que deseen predecir su puntaje en las pruebas Saber 11. Las gráficas seleccionadas son:

- Comparación del puntaje promedio por departamento
- Comparación del puntaje promedio por ubicación del colegio y por tipo de colegio
- Comparación del puntaje promedio por estrato en que vive el estudiante

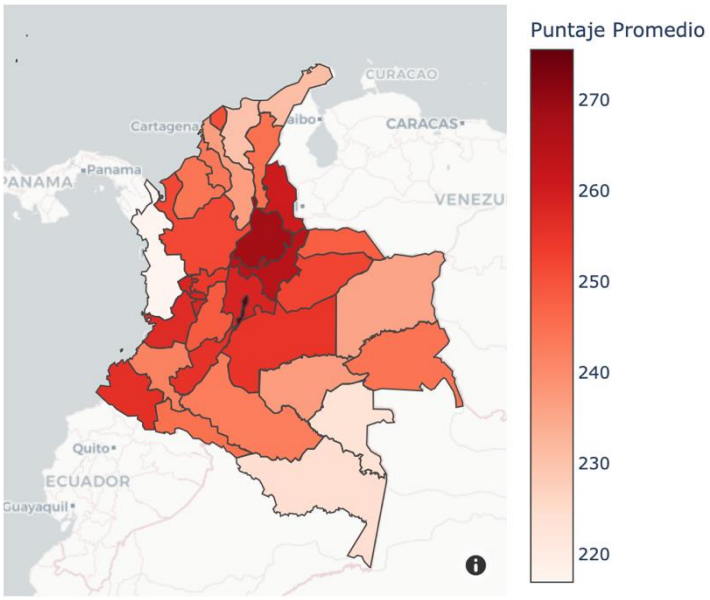

Figura 2. Puntaje Promedio por Departamento

Al analizar los resultados promedio de la prueba Saber 11 por departamento, se observa una amplia variabilidad en los puntajes promedio entre los departamentos. La ciudad de Bogotá D.C. obtiene el puntaje promedio más alto con 275.54, mientras que Chocó registra el puntaje más bajo con 216.98. Esta diferencia de casi 60 puntos indica una disparidad significativa en el rendimiento académico entre estas regiones. También podemos observar que la capital del país presenta un puntaje promedio por encima de la media nacional (252.28), lo que sugiere una buena calidad educativa en la ciudad.

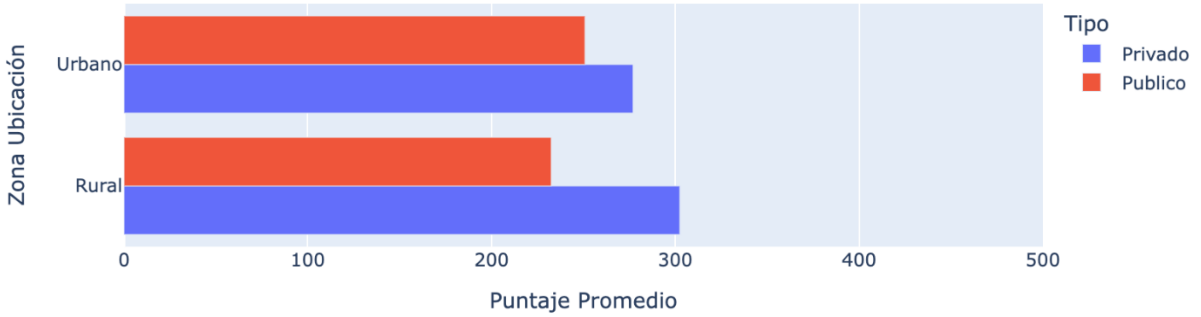

Figura 3. Puntaje Promedio por Zona y por Tipo

Los resultados obtenidos de la prueba Saber 11 muestran diferencias significativas en los promedios según la ubicación geográfica de los colegios y el tipo de institución. Al analizar los datos, se observa que, en las zonas rurales, los colegios privados obtuvieron un promedio de 302.45 puntos, mientras que los colegios públicos alcanzaron un promedio de 232.52 puntos. Esta disparidad sugiere que los colegios privados en áreas rurales tienden a superar ampliamente a los colegios públicos en términos de resultados académicos. Por otro lado, en las zonas urbanas también se identificaron diferencias notables. Los colegios privados en entornos urbanos obtuvieron un promedio de 277.04 puntos, mientras que los colegios públicos alcanzaron un promedio ligeramente inferior de 250.85 puntos. Aunque la brecha no es tan amplia como en las zonas rurales, aún existe una diferencia significativa entre los resultados de los colegios privados y públicos en áreas urbanas. Estos hallazgos indican que, en general, los colegios privados presentan un desempeño académico superior en comparación con los colegios públicos en ambas zonas, tanto rurales como urbanas. Estos resultados pueden atribuirse a una

variedad de factores, como mayores recursos, infraestructuras más adecuadas, docentes más calificados o metodologías pedagógicas diferenciadas.

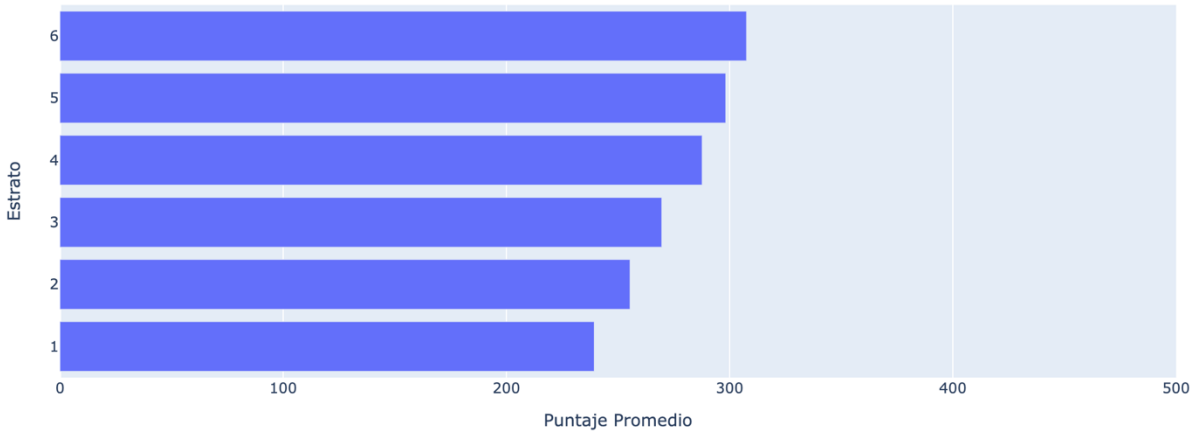

Figura 4. Puntaje Promedio por Estrato

Se realizaron análisis de los resultados de la prueba Saber 11 en relación con el estrato socioeconómico de los estudiantes. Los hallazgos revelaron que existe una marcada diferencia en los promedios de resultados entre los estratos. El estrato 6 obtuvo el promedio más alto con 307.55 puntos, seguido de cerca por el estrato 5 con 298.23 puntos. A medida que se desciende en los estratos, se observa una disminución gradual en los promedios. El estrato 4 obtuvo un promedio de 287.67 puntos, seguido por el estrato 3 con 269.59 puntos. Los estratos 2 y 1 presentaron los promedios más bajos con 255.38 y 239.34 puntos, respectivamente.

Estos resultados resaltan una clara disparidad en los desempeños académicos según el estrato socioeconómico. Es evidente que los estudiantes pertenecientes a estratos más altos tienen un rendimiento académico superior en comparación con aquellos de estratos más bajos. Esta brecha puede ser atribuida a una serie de factores, incluyendo diferencias en los recursos educativos disponibles, acceso a oportunidades de aprendizaje adicionales, apoyo e influencia familiar.

5 Modelo de Red Bayesiana

Para la construcción del modelo de Red Bayesiana se realizó la estimación de la estructura por medio de dos métodos: por medio del puntaje BIC y el puntaje K2. Mirando cuales eran los arcos y nodos que ambos métodos estimaban junto con lo aprendido en el EDA se obtuvo la Red Bayesiana a utilizar. El grafo final cuenta con 11 nodos y 16 arcos dirigidos que permiten representar el problema por medio de una red bayesiana, que permita predecir el puntaje esperado por el estudiante.

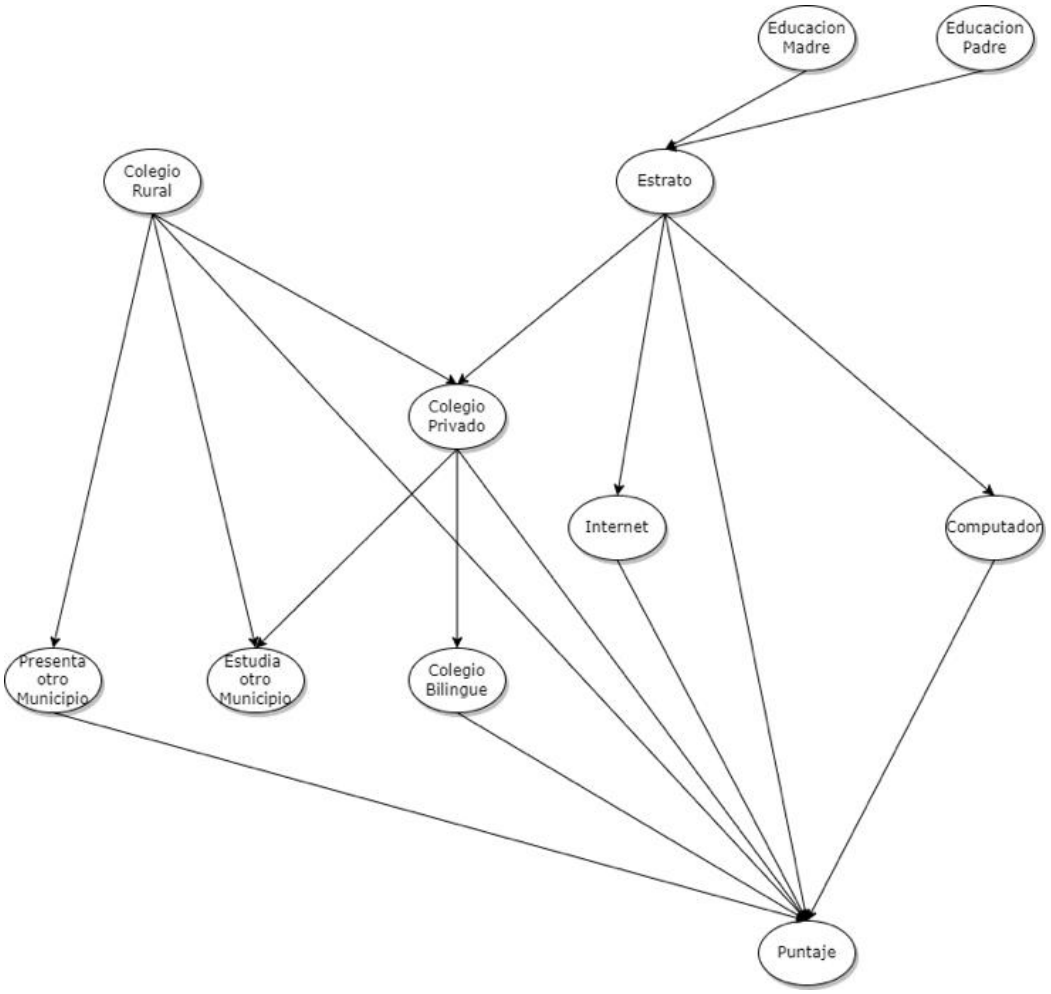

Figura 5. Grafo

6 Entrenamiento del modelo

Como se cuenta con una cantidad significativa de datos, más de 3 millones de datos, para evaluar el modelo se utiliza la metodología de usar un train set y un test set. Se utilizan para los porcentajes de cada set la *Rule of Thumb* de la industria de entrenar el modelo con el 75% de los datos y probarlo con el 25% restante, seleccionados de manera aleatoria. Como se observa en la gráfica:

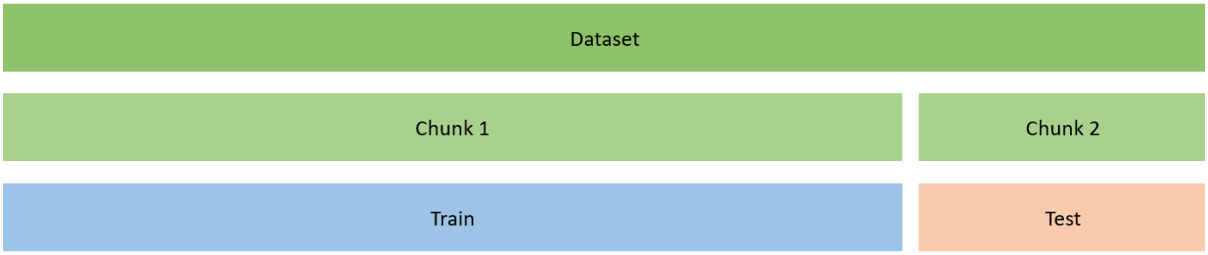

Figura 6. Train y Test Set

7 Testing

Dado que se está trabajando con una Red Bayesiana Discreta las variables del modelo deben ser discretas. Ante esta situación, nuestra variable a predecir, el puntaje global, es continua entre 0 y 500. Por lo que se tomó la decisión de discretizar los datos de puntaje global por medio de

intervalos de 50 puntos, es decir, se transforma en una variable discreta con 10 posibles valores, como se muestra a continuación:

Rango Puntaje	Grupo
[0 – 50]	1
[50 – 100]	2
[100 – 150]	3
[150 – 200]	4
[200 – 250]	5
[250 – 300]	6
[300 – 350]	7
[350 – 400]	8
[400 – 450]	9
[450 – 500]	10

Al calcular la evidencia par cada dato se obtiene una lista de probabilidades, que corresponde a cada categoría de la variable. Por lo que aquella categoría con mayor probabilidad se clasificará como que el estudiante pertenece a dicha categoría.

8 Evaluación del Modelo

Al realizar la clasificación para el test set, es posible comparar los valores reales de los estudiantes con los valores predichos. El 25% de los datos, es decir el test set tiene 823,992 filas, por lo que la matriz de confusión se ve de la siguiente forma:

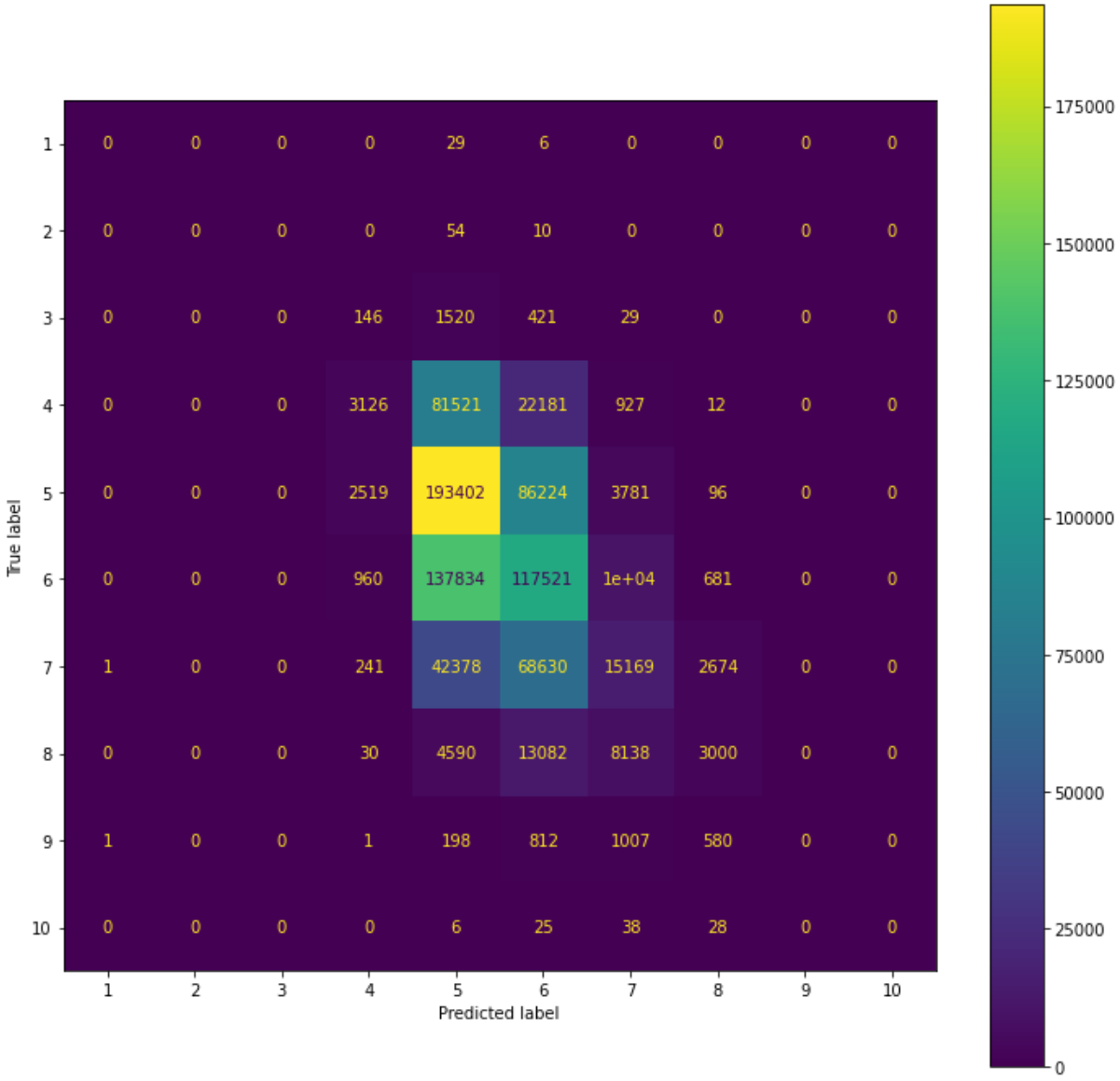

Figura 7. Matriz de Confusión 10 Categorías

Ahora con esta matriz de confusión calculemos el precision, el recall y el F1-Score.

Categoria	precision	recall	f1-score	Numero
1	0	0	0	35
2	0	0	0	64
3	0	0	0	2116
4	0.45	0.03	0.05	107767
5	0.42	0.68	0.52	286022
6	0.38	0.44	0.41	267359
7	0.38	0.12	0.18	129093
8	0.42	0.1	0.17	28840
9	0	0	0	2599
10	0	0	0	97
Accuracy			0.69	
Avg	0.21	0.14	0.13	
Weighted Avg	0.40	0.41	0.35	

Como se puede observar el modelo no es muy bueno, pocas veces se supera el 0.5 de precision y de recall.

Ante esta situación se decide probar con menos categorías, con intervalos de 100 puntos, es decir, 5 categorías.

Rango Puntaje	Grupo
[0 – 100]	1
[100 – 200]	2
[200 – 300]	3
[300 – 400]	4
[400 – 500]	5

Repetimos el procedimiento anterior

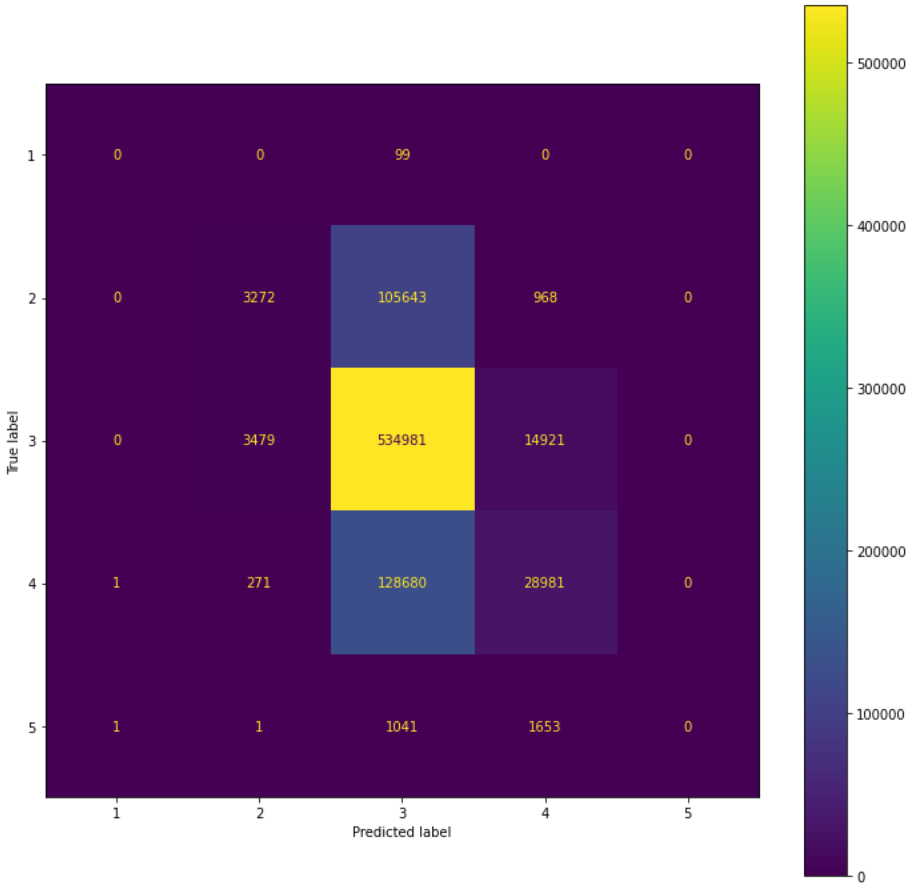

Figura 8. Matriz de Confusión 5 Categorías

Categoria	precision	recall	f1-score	Numero
1	0	0	0	99
2	0.47	0.03	0.06	109883
3	0.69	0.97	0.81	553381
4	0.62	0.18	0.28	157933
5	0	0	0	2696
Accuracy			0.40	
Avg	0.36	0.24	0.23	
Weighted Avg	0.64	0.69	0.61	

Podemos ver que con menos categorías aumenta drásticamente el desempeño del modelo, siendo ahora mejor que un modelo aleatorio. Por lo que, aunque perdemos capacidad de inferencia, ya que ahora los intervalos son de 100 puntos, podemos obtener mejores medidas de desempeño para el modelo. De nada sirve que se clasifique en categorías más precisas o exactas si el modelo no es preciso en sí.

Por lo tanto, se decide que el mejor modelo es aquel con 5 categorías y este será el usado para el tablero en Dash que será usado por los estudiantes para predecir su puntaje.