Daniel Forero Olano 201924895

Juan Fernando Martinez 201533234

**Reporte Proyecto 3**

**Extracción y exploración de datos**

Lo primero que quisimos hacer en este punto fue determinar cuántos datos de cada tema se encontraban

|  |  |
| --- | --- |
| Tema | Cantidad de columnas |
| Estudiante | 17 |
| Familia | 9 |
| Colegio | 17 |
| Periodo | 1 |
| Desempeño y puntaje | 7 |

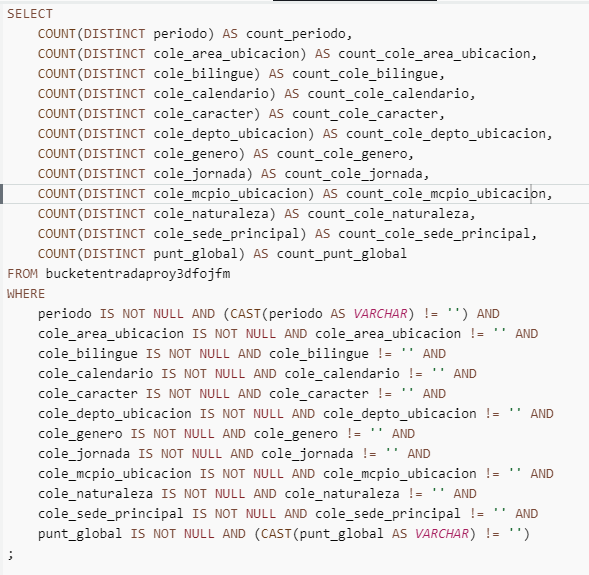
Después de verificar esto notamos que los temas con más información son el estudiante y el colegio por lo cual debíamos seleccionar alguno de los dos y nos decantamos por enfocarnos en los datos del colegio agregando el periodo y el puntaje global para tener una columna objetivo y una métrica temporal de comparación.

Con esto ya definido igualmente seguíamos incluyendo muchos datos (más de 7 millones) por lo cual empezamos a filtrar con nuestra primera consulta

select distinct periodo FROM bucketentradaproy3dfojfm order by periodo desc

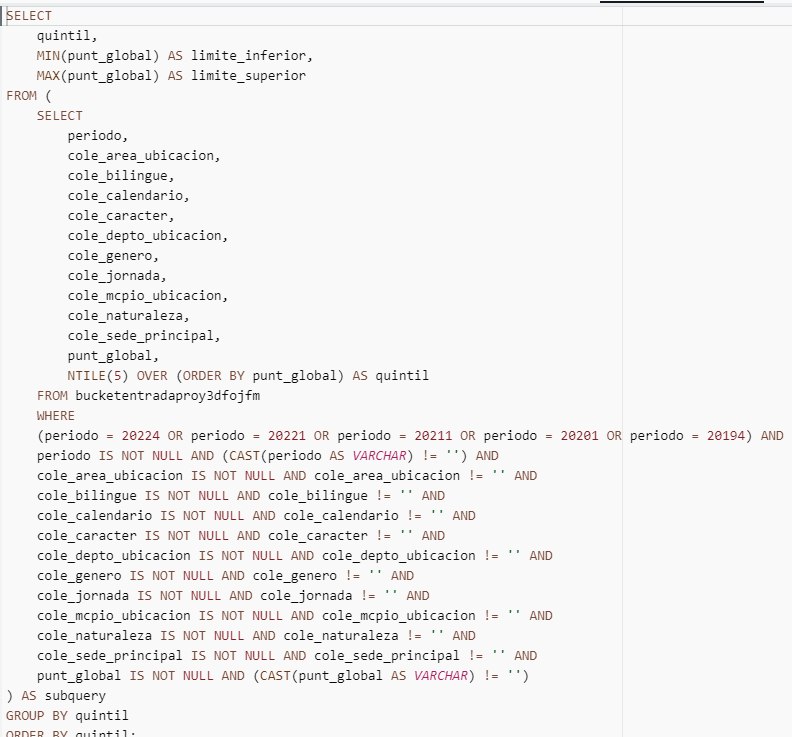
Con la cual entonces nos quedamos con los últimos 5 periodos (20224, 20221, 20211, 20201, 20194) y se reducen

Posteriormente ya con los datos que queremos seleccionar hacemos la siguiente consulta omitiendo vacíos para ver cuántos datos diferentes tenemos en cada columna

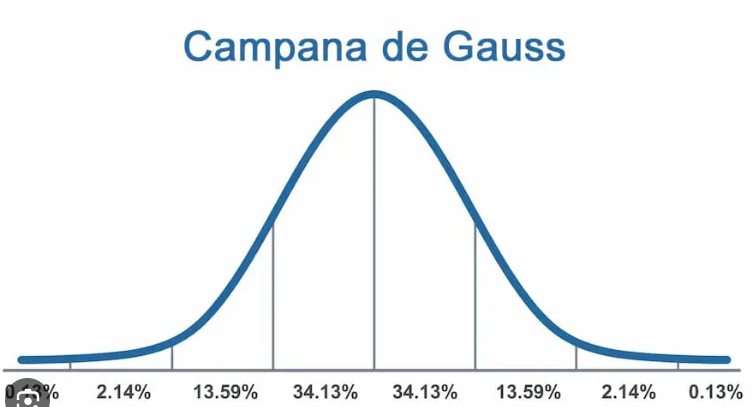


Y dentro de los resultados los dos más inquietantes son el del municipio de ubicación del colegio que tiene 1396 valores diferentes y el de puntaje global el cual tiene 484 valores diferentes por lo cual tomamos dos determinaciones la primera no usar la información del municipio pues ya tenemos la del departamento y la segunda es discretizar el puntaje en grupos.

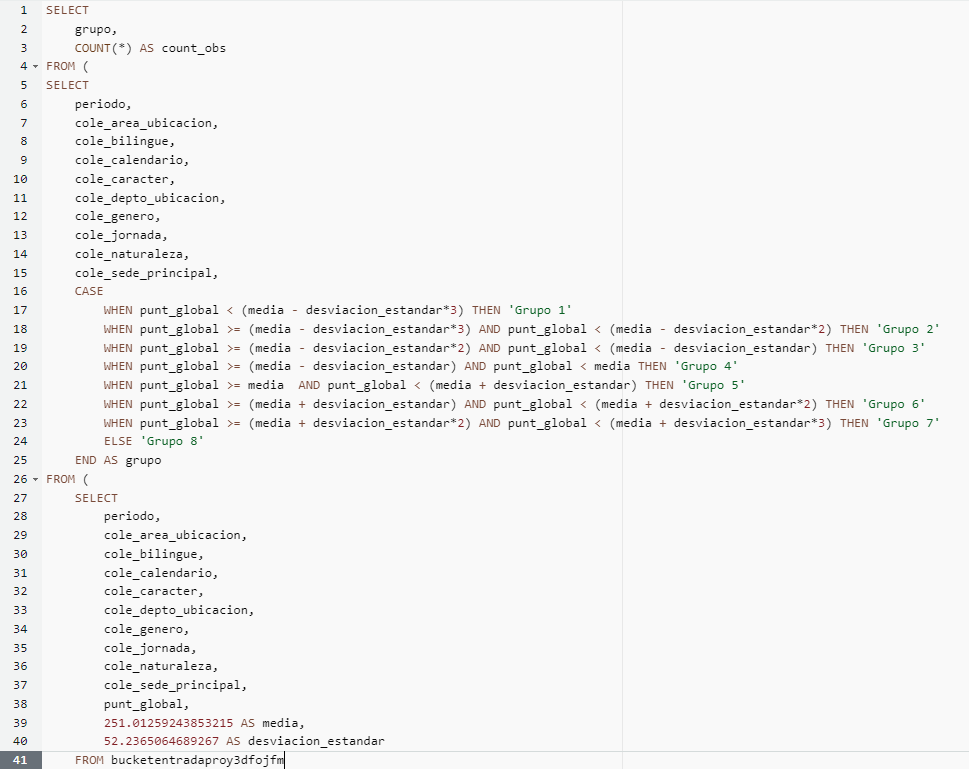
Nuestra primera idea fue dividirlo en quintiles el puntaje, pero usando este query



Nos dimos cuenta que el ultimo quintil iba de 298 a 495 por lo cual la información resultante no era muy provechos y explicativa, entonces nos plantemos dividir la población como una campana de Gauss



Para lograr esto entonces primero necesitábamos calcular la media y la desviación las cuales son funciones no aceptadas en Athena por lo cual tuvimos que descargar una primera versión de nuestros datos con los filtros de periodo y vacíos para obtener estos valores en Python, los cuales fueron una media de 251.01259243853215 y desviación estándar de 52.2365064689267. Ya con estos datos se plantearon los grupos y con el siguiente query obtuvimos la información

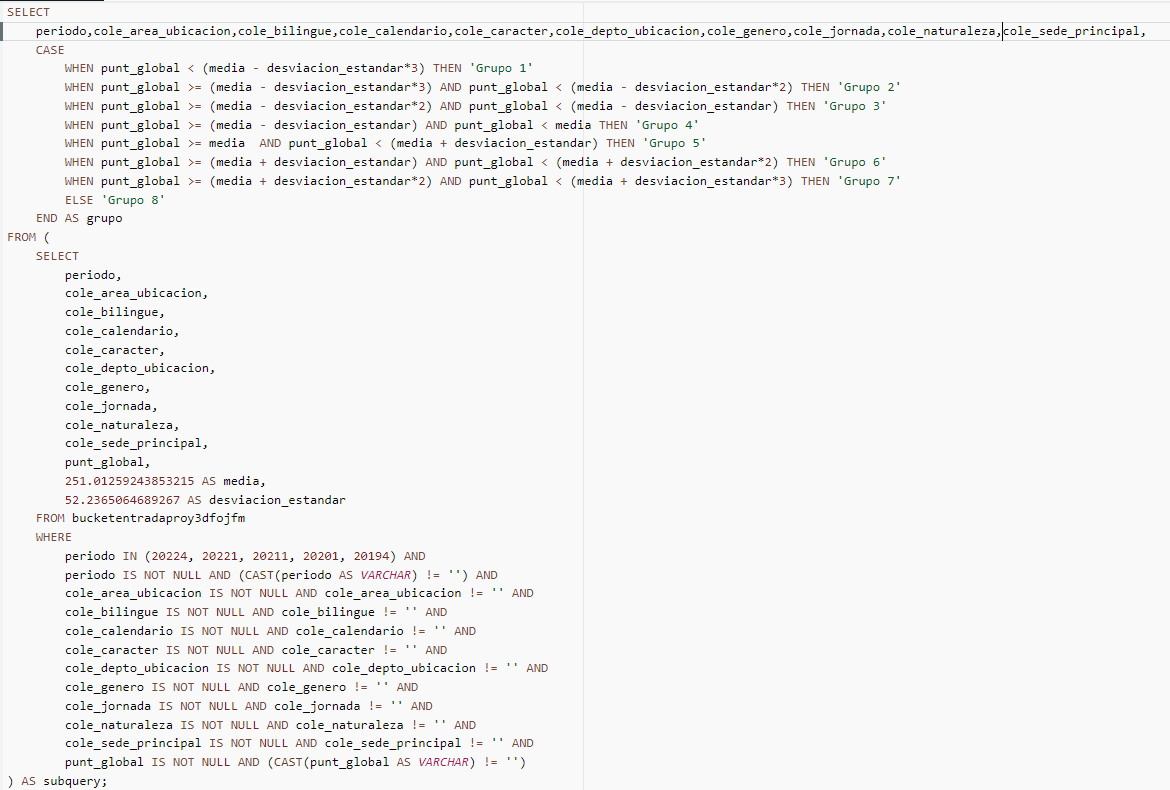


Texto, Carta

Descripción generada automáticamente

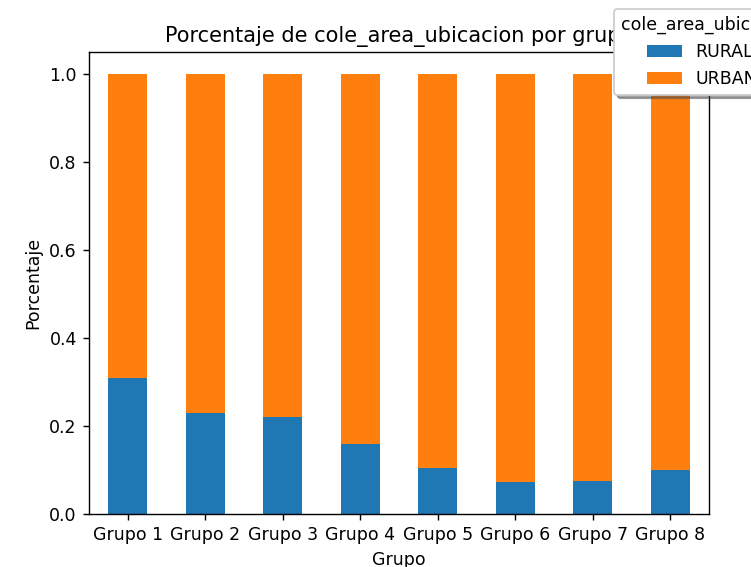
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Grupo** | **Lim. Inf [** | **Lim. Sup )** | **Observaciones** | **%** | **% teórico** |
| **1** | 0 | 94 | 524 | 0.03% | 0.13% |
| **2** | 94 | 147 | 7,496 | 0.41% | 2.14% |
| **3** | 147 | 199 | 314,551 | 17.17% | 13.59% |
| **4** | 199 | 251 | 635,167 | 34.67% | 34.13% |
| **5** | 251 | 303 | 556,623 | 30.39% | 34.13% |
| **6** | 303 | 355 | 271,533 | 14.82% | 13.59% |
| **7** | 355 | 408 | 43,827 | 2.39% | 2.14% |
| **8** | 408 | 500 | 2,172 | 0.12% | 0.13% |
|  |  |  | 1,831,893 |  |  |

Con esta información ya obtenida importamos los datos como csv para trabajarlos en Python con esta consulta



Con lo cual ya entonces tenemos nuestra base de 11 columnas y 1,831,893 registros con la que revisamos la relación entre ellas, dentro de las más destacadas están

Área de ubicación del colegio:



Acá podemos notar como los mejores grupos tienen un mayor porcentaje de participación de colegios en áreas urbanas lo que nos da una pista sobre su influencia

Si el colegio es bilingüe:

Gráfico, Gráfico de barras

Descripción generada automáticamente

Acá podemos notar como los mejores grupos tienen un mayor porcentaje de participación de colegios bilingües lo que indica que puede ser una variable significativa para determinar el objetivo

El carácter del colegio:

Gráfico, Gráfico de barras

Descripción generada automáticamente

Por último, podemos notar como los mejores grupos tienen un mayor porcentaje de participación de colegios de carácter académico implicando una relación relevante entre esta variable y la variable objetivo

**Definición del producto**

Ya con los datos analizados y escogidos entonces se define que el cliente de este producto es el gobierno nacional para poder evaluar cuales son las mejores condiciones de educación posibles que permitan un mejor desempeño en las pruebas estatales. Con esto en mente se procede a mirar posibles modelos de interés para clasificar el rendimiento.

**Modelos**

Para proponer un modelo relevante para el objetivo, estimamos una estructura a partir de los datos por puntaje, usando dos métodos k2 y bic, e iterando sobre los modelos obtenidos para eliminar los nodos que no son significativos sobre la variable de interés. Como datos de entrenamiento se tomó una muestra del 20% de los datos, y la variable objetivo el grupo del puntaje.

Aquí presentamos las iteraciones y sus puntajes.

**K2 scoring:**

|  |  |
| --- | --- |
| **Iteración 1** | **Iteración 2** |
|  |  |
| K2 score: -12019458.39 | K2 score: -10676158.92 |

**BIC Scoring:**

|  |  |
| --- | --- |
| **Iteración 1** | **Iteración 2** |
|  |  |
| BIC score: -12070778.65 | BIC score: -10484563.90 |
| **Iteración 3** |  |
|  |  |
| BIC score: -10397566.45 |  |

Según esto, inicialmente elegimos el modelo con el puntaje más cercano a 0, que es el de la tercera iteración del modelo obtenido por puntaje BIC además de ser el menos complejo, pero para estar seguros evaluaremos su desempeño frente a la última iteración del modelo dado por K2 scoring.

**Evaluación de los mejores modelos:**

Lo primero es comparar dentro del modelo dado por BIC si funciona mejor un estimador bayesiano con distribución a priori uniforme o con estimador por máxima verosimilitud

Gráfico, Gráfico de barras

Descripción generada automáticamenteGráfico, Gráfico de barras

Descripción generada automáticamente

Como se puede observar el rosado que es por estimación bayesiana y el verde que es por máxima verosimilitud dan igual en accuracy y f1 por lo tanto por tiempo de ejecución nos quedamos entre estos dos con el modelo de estimación por máxima verosimilitud.

Posteriormente se repite el mismo ejercicio para el modelo dado por K2 Scoring

Gráfico

Descripción generada automáticamenteGráfico

Descripción generada automáticamente

Igual que con el modelo dado por Bic se puede observar el azul que es por estimación bayesiana y el verde que es por máxima verosimilitud dan igual en accuracy y f1 por lo tanto por tiempo de ejecución nos quedamos entre estos dos con el modelo de estimación por máxima verosimilitud

Por último, comparando el modelo dado por Bic con estimación por máxima verosimilitud contra el modelo dado por K2 con estimación por máxima verosimilitud se obtiene lo siguiente:

Gráfico, Gráfico de barras

Descripción generada automáticamenteGráfico

Descripción generada automáticamente

Como se puede observar en este caso el que mejor se comporta es el modelo dado por K2 scoring (verde claro), por lo cual es el modelo en el que se basara el producto.

**Análisis modelo final:**

|  |  |
| --- | --- |
| Métrica | Valor |
| Accuracy | 0.4041170481932643 |
| F1\_Macro | 0.1899680406181743 |
| F1\_Micro | 0.4041170481932643 |

El modelo de clasificación exhibe un rendimiento aceptable, logrando una precisión del 40.41% mucho mejor que un aleatorio el cual tendría 12.5% dado que son 8 posibles respuestas. Aunque las puntuaciones F1-Macro y F1-Micro son moderadas, con valores de 18.99% y 40.41%, respectivamente, sugieren que el modelo es capaz de realizar predicciones acertadas en un porcentaje significativo de casos. Este desempeño positivo proporciona una base sólida para futuras mejoras y ajustes, demostrando el potencial del modelo para realizar predicciones más confiables que la aleatoriedad.