Sebastián Arango Crispín – 201921816
Sofia Buitrago Carvajal – 202020742
Tomás Liévano Casas – 201920980

Universidad de los Andes
Analítica Computacional IIND-4130

Proyecto 3 – Pruebas Saber 11

1. Preguntas de negocio y plan de acción

En primer lugar, se escogió como cliente Secretarías de educación de alcaldías y gobernaciones, interesadas en conocer cómo sus estudiantes e instituciones se comparan entre sí y cómo se comparan con las de otros municipios y departamentos. Además, nos enfocaremos en los resultados obtenidos en el departamento de Antioquia y las principales preguntas de negocio que se buscan resolver son:

1. ¿Qué factores son los que más influyen en el resultado de un estudiante en las pruebas Saber 11?
2. ¿En qué ámbitos se deben enfocar las alcaldías y gobernaciones para mejorar los resultados del examen en el departamento?

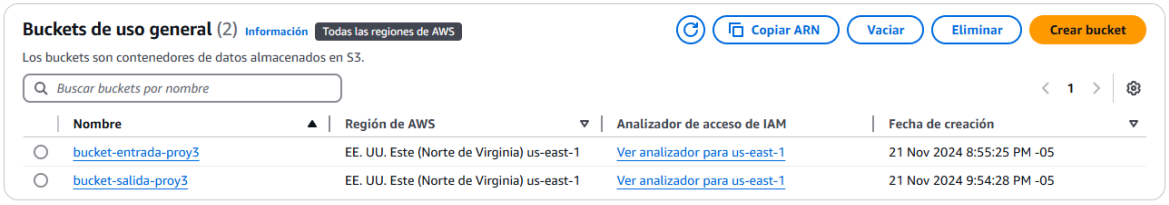
Estas preguntas se abordarán mediante visualizaciones como histogramas para analizar la distribución de variables clave en relación con el puntaje obtenido, gráficos de barras para variables categóricas y box plots para ver qué factores son los que más influyen en el resultado. Además, se usará un modelo predictivo de clasificación basado en redes neuronales para identificar patrones que contribuyan a predecir la mejora o empeoramiento de los resultados del examen. Este análisis permitirá tomar decisiones informadas que permitan crear o enfocar campañas educativas y así mejorar el aprendizaje de los estudiantes del departamento.

Analista de negocios: **Tomás Liévano.**

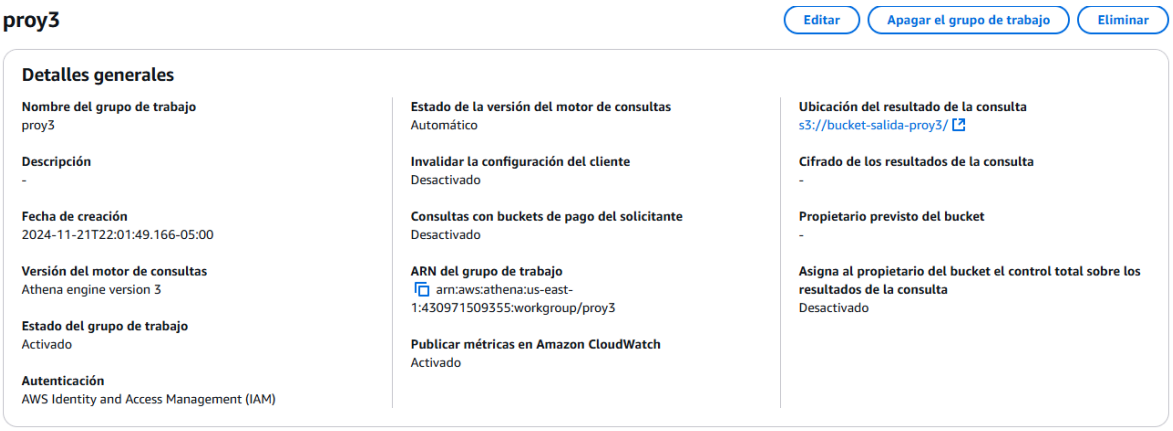
2. Datos

2.1 Limpieza y alistamiento de datos

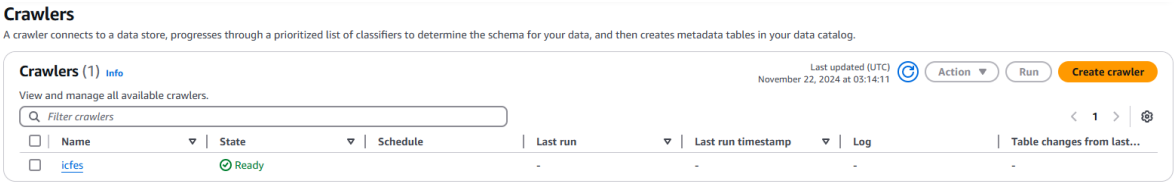
Creación de buckets en AWS S3 con los datos del ICFES:



Creación del grupo de trabajo:



Crawler:



A partir de una revisión inicial de los datos, se filtró en primer lugar por el departamento asignado, de manera que se tomaron solamente los valores correspondientes a Antioquia como departamento en el que se tomó la prueba. Esto permitió reducir el número de observaciones de 7,110,000 a 1,009,317.

Ahora bien, se evaluó la significancia de cada una de las 51 variables inicialmente presentadas en los datos con respecto a las preguntas de interés definidas en el análisis de negocio. Así pues, se decidió quitar aquellas variables que no servían para caracterizar comportamientos o que contuvieran información redundante, como por ejemplo una variable para el código de departamento de residencia y el nombre de este mismo. En estos casos, se dejó solo los nombres, pues son más informativos para los fines del proyecto. De esta manera, se eliminaron las siguientes variables:

- 'COLE_COD_MCPIO_UBICACION',
- 'ESTU_COD_RESIDE_MCPIO',
- 'PERIODO',
- 'ESTU_COD_MCPIO_PRESENTACION',
- 'COLE_CODIGO_ICFES',
- 'ESTU_ESTUDIANTE',
- 'COLE_COD_DEPTO_UBICACION',
- 'COLE_COD_DANE_ESTABLECIMIENTO',
- 'ESTU_TIPODOCUMENTO',
- 'ESTU_COD_RESIDE_DEPTO',
- 'ESTU_CONSECUTIVO',
- 'ESTU_COD_DEPTO_PRESENTACION',
- 'DESEMP_INGLES',
- 'COLE_COD_DANE_SEDE',
- 'COLE_DEPTO_UBICACION'

Así, se obtienen finalmente 36 variables. Ahora bien, se identificaron numerosos valores vacíos para diversas variables. Por esta razón, se definió que para las variables que tienen un porcentaje menor o igual al 1% de valores vacíos, se iban a eliminar estas observaciones. No obstante, para aquellos que tuvieran un valor mayor, se realizó una imputación con base en la moda de los valores de la columna respectiva También se eliminaron aquellas observaciones donde la variable de salida 'PUNT_GLOBAL' tiene valores faltantes, pues no son de utilidad para el análisis ni para el modelo de predicción.

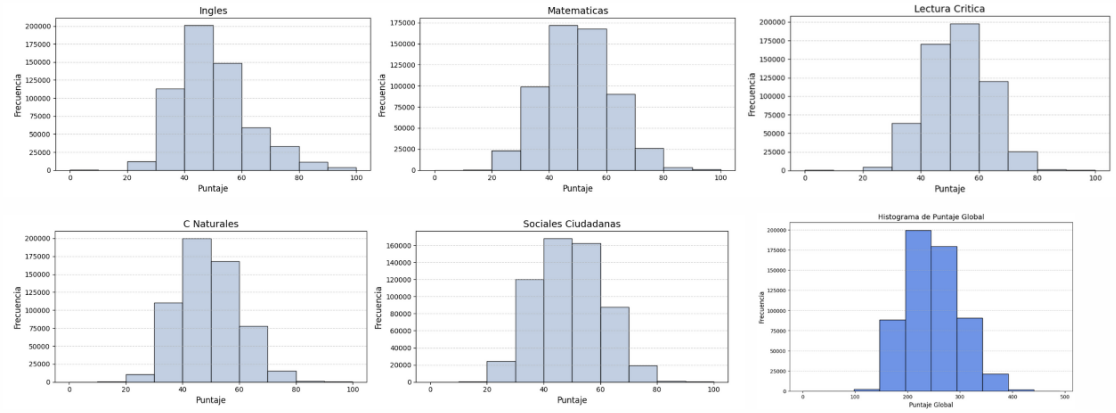
Ingeniería de datos: **Sebastián Arango Crispín.**

2.2 Exploración de datos

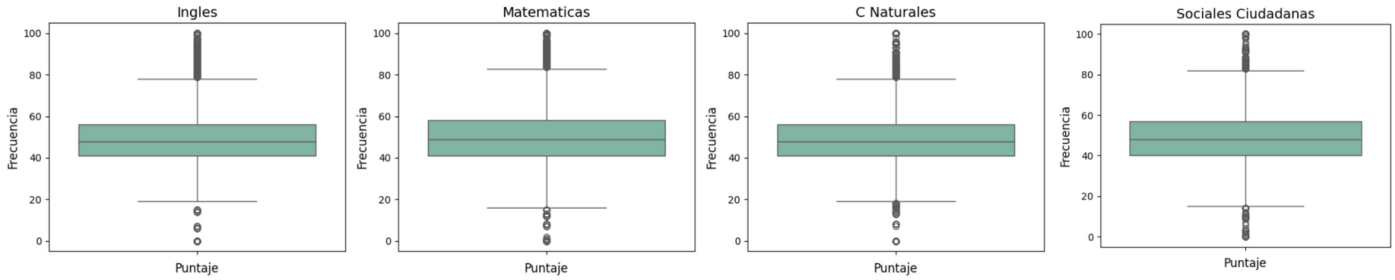
Dado que el cliente objetivo es el área de educación de la Gobernación de Antioquia, se identificó un posible interés en explorar los factores asociados con las características de los colegios, las condiciones sociodemográficas de los estudiantes y el desempeño en las diferentes materias evaluadas en las pruebas Saber 11. Por ello, el análisis de las estadísticas descriptivas se enfocó en dichas variables.

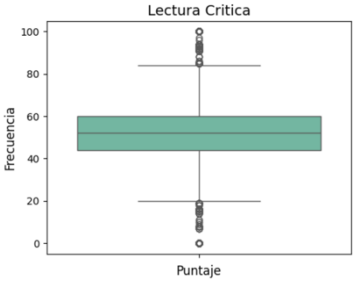
Sebastián Arango Crispín – 201921816
Sofia Buitrago Carvajal – 202020742
Tomás Liévano Casas – 201920980

Universidad de los Andes
Analítica Computacional IIND-4130

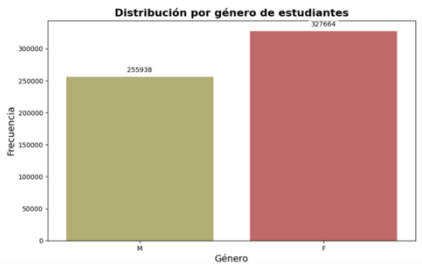


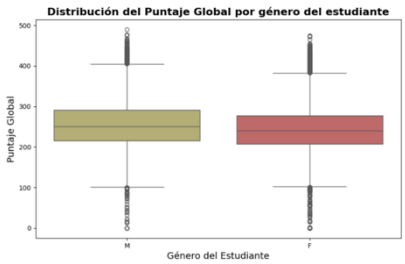
Se calcularon las medidas de tendencia central y de dispersión para las variables continuas, incluyendo los puntajes en inglés, matemáticas, ciencias naturales, competencias ciudadanas, lectura crítica y el puntaje global. Se evidenció que el promedio para cada materia estaba entre 48.46 y 52.06 y se tiene un puntaje mínimo de 0 y máximo de 100 para estas mismas materias. En cuanto al puntaje global se encontró un puntaje mínimo de 0, un máximo de 490 y un promedio de 248.38. Adicionalmente, como se evidencia anteriormente, los histogramas muestran que los puntajes de distribuyen alrededor de la media pero también presentan valores que se desvían hacia ambos extremos, tanto a la izquierda como a la derecha.



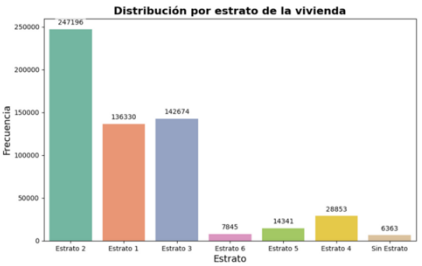


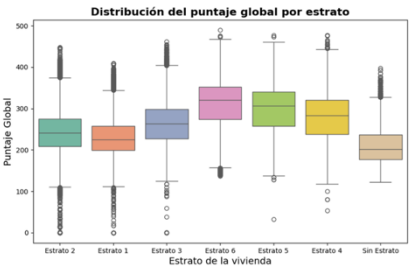
Por otro lado, se calcularon las frecuencias de los valores únicos presentes en las variables categóricas para identificar patrones de distribución. Posteriormente, estas variables fueron analizadas mediante gráficos de barras y diagramas de caja, los cuales permitieron visualizar tanto la proporción de registros en cada categoría como la variabilidad de los valores asociados a cada clasificación. Este enfoque facilitó la comparación entre categorías y la identificación de tendencias relevantes en los datos.





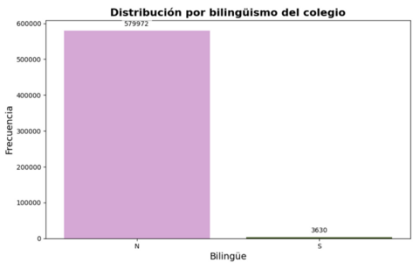
En la variable **género del estudiante** se encontró que se registraron más mujeres (327664) en las pruebas que hombres (255938). En el diagrama de caja se puede observar que el promedio y los cuartiles del puntaje global por género del estudiante son similares para las dos categorías.

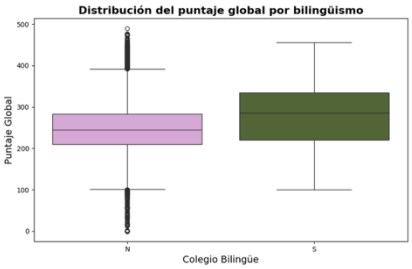




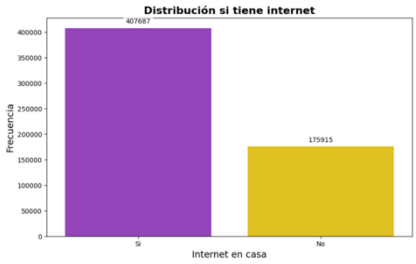
En la variable **estrato de la vivienda** se puede evidenciar que la mayoría de registros son para los estratos 1, 2 y 3. El análisis con boxplots, considerando el total por categoría, muestra que la media y los cuartiles del puntaje global tienden a incrementarse conforme aumenta el estrato. Además, los registros correspondientes a viviendas sin estrato presentan los puntajes globales más bajos.

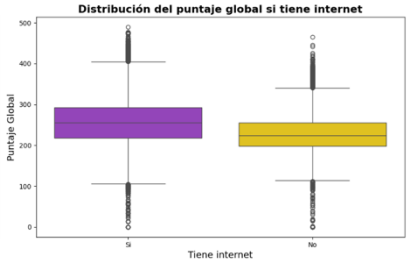
Sebastián Arango Crispín – 201921816
Sofia Buitrago Carvajal – 202020742
Tomás Liévano Casas – 201920980



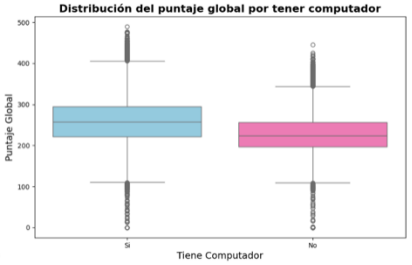


En la variable **bilingüismo del colegio** se observa que la cantidad de colegios bilingües (3630) es baja. El boxplot revela que los cuartiles de los puntajes son más altos para estos colegios; sin embargo, es importante considerar la diferencia en la cantidad de observaciones entre las categorías.

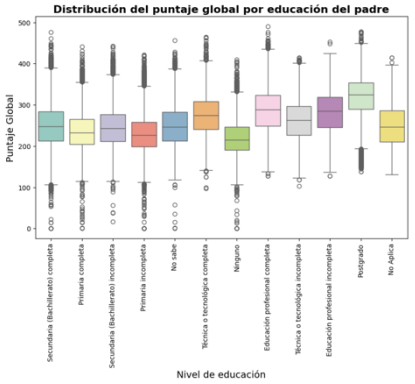


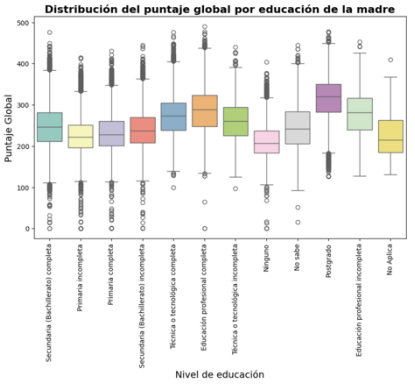


En la variable **si tiene internet** hay 407687 registros de los estudiantes que sí cuentan con internet en casa, mientras que, 175915 no lo tienen. En el boxplot se puede observar que los percentiles del 25% al 100% para el puntaje global son superiores para los registros que sí tienen internet en casa.



En la variable **si tiene computador** hay 380751 registros de los estudiantes que sí cuentan con un computador en casa, mientras que, 202851 no lo tienen. En el boxplot se puede observar que los percentiles del 25% al 100% para el puntaje global son superiores para los registros que sí tienen internet en casa.





En la variable correspondiente al **nivel educativo de la madre**, las tres categorías con mayor número de registros son "secundaria completa", "primaria incompleta" y "secundaria incompleta". En contraste, "posgrado" cuenta con 11,932 registros y "técnica completa" con 59,263. Según el boxplot, los puntajes globales más altos se encuentran en las categorías "posgrado" y "educación profesional completa e incompleta", mientras que las categorías "educación primaria completa e incompleta" y "ninguno" registran los puntajes más bajos. En cuanto al **nivel educativo del padre**, las categorías con más registros coinciden parcialmente con las de la madre: "secundaria completa", "primaria incompleta" y, en tercer lugar, "secundaria incompleta". La categoría "posgrado" tiene 11,609 registros, y "técnica completa" alcanza 36,727. Los boxplots muestran que las categorías con mayores percentiles en puntaje global son similares a las observadas en el caso de la madre, y los valores más bajos corresponden a la categoría "ninguno".

Analista de datos: **Sofía Buitrago Carvajal.**

3. Modelos

3.1 Modelamiento

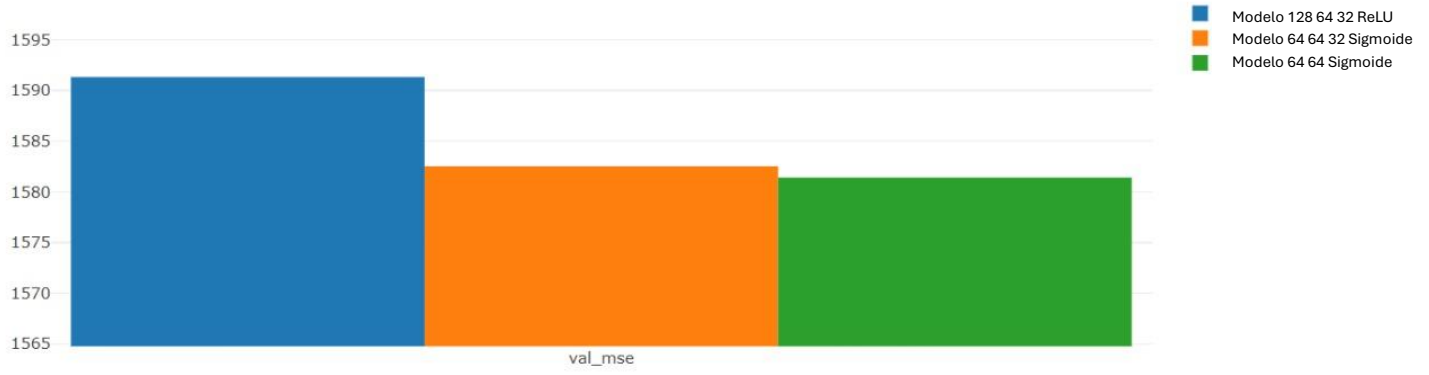
Considerando las preguntas de interés, se propone hacer un modelo de redes neuronales que permita predecir el puntaje global que va a recibir un estudiante con base en algunas de las características de sus condiciones socio económicas y educacionales. Así, en primer lugar, se decidió quitar todas las variables presentes que fueran muy específicas de cada presentante del examen. En otras palabras, aquellas que no permitieran realizar una generalización de las tendencias. Adicionalmente, a pesar de que la edad sería una variable importante e incluir en el modelo, debido a que solo se cuenta con la fecha de nacimiento y no se sabe la fecha de presentación del examen sino solo un periodo

general, entonces se decidió excluir estas variables del modelo. Además, se eliminaron las variables correspondientes a los resultados de cada prueba específica (inglés, matemáticas, sociales, etc.), pues el puntaje global se calcula directamente a partir de estas y no tiene sentido predecirlo con ellas como variables si tienen un cálculo directo. Entonces, las 20 variables que se decidió dejar para entrenar el modelo fueron las siguientes:

- **COLE_AREA_UBICACION:** Ubicación de la sede
- **COLE_BILINGUE:** ¿Es colegio bilingüe?
- **COLE_CALENDARIO:** Calendario del Establecimiento
- **COLE_CARACTER:** Carácter del establecimiento
- **COLE_GENERO:** Genero del Establecimiento
- **COLE_JORNADA:** Jornada de la Sede
- **COLE_NATURALEZA:** Naturaleza del establecimiento
- **COLE_SEDE_PRINCIPAL:** ¿Es la sede Principal?
- **ESTU_ESTADOINVESTIGACION:** ¿Permite usar sus datos para Investigaciones?
- **ESTU_GENERO:** Genero del examinando
- **ESTU_PRIVADO_LIBERTAD:** ¿Es privado de la libertad?
- **FAMI_CUARTOSHOGAR:** ¿Cuántos cuartos tiene su hogar?
- **FAMI_EDUCACIONMADRE:** Nivel de estudios de la madre
- **FAMI_EDUCACIONPADRE:** Nivel de estudios del padre
- **FAMI_ESTRATOVIVIENDA:** Estrato del examinando
- **FAMI_PERSONASHOGAR:** ¿Con cuantas personas vive?
- **FAMI_TIENEAUTOMOVIL:** ¿Tiene automóvil?
- **FAMI_TIENECOMPUTADOR:** ¿Tiene computador?
- **FAMI_TIENEINTERNET:** ¿Tiene internet?
- **FAMI_TIENELAVADORA:** ¿Tiene lavadora?

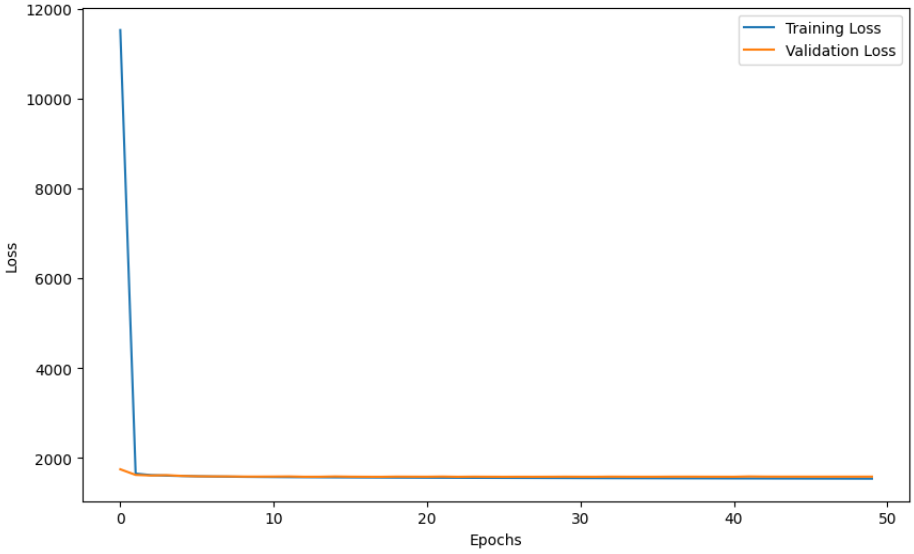
Así, ya que todas las variables de entrada son categóricas, se procedió a hacer una dumificación de estas para poder ingresarlas como inputs en el modelo de redes neuronales. Debido a que es un número elevado de variables, y teniendo en cuenta también que estas variables tienen numerosas categorías cada una, entonces el número de variables dumificadas con las que se termina trabajando es de 98. Por esta razón, se decide usar la herramienta de One Hot Encoder, que permite guardar la configuración en la que se hace la dumificación. Esto permite replicabilidad, el igual que permite cargar la configuración a la hora de cargar el modelo y hacer la predicción.

Posteriormente, se evaluaron diferentes modelos con el fin de ver cual predecía mejor. De esta manera, se evaluaron 3 modelos diferentes. El primer modelo que se evaluó contaba con 3 capas densas, la primera con 128 neuronas, la segunda con 64 neuronas y la tercera con 32 neuronas. Además, tenía una función de activación ReLU. El segundo modelo que se evaluó contaba también con 3 capas densas, esta vez con 64 neuronas en las primeras dos capas y 32 neuronas en la tercera. Igualmente, se usó una función de activación sigmoide. Finalmente, el tercer modelo que se evaluó contaba con dos capas densas, ambas con 64 neuronas, y función de activación sigmoide. Las siguientes tabla y figura muestran la comparación de los 3 modelos con respecto a su MSE, obtenida a partir de MLFlow.



Modelo 128 64 32 ReLU	Modelo 64 64 32 Sigmoide	Modelo 64 64 Sigmoide
1591.3	1582.5	1581.4

Como se puede ver, el tercer modelo fue el que obtuvo un valor menor de MSE, por lo que se eligió este para realizar la predicción de variables. Además, en el momento del entrenamiento, se graficó la comparación entre el valor de la función de pérdida para los datos de entrenamiento y para los datos de validación con respecto a las épocas.



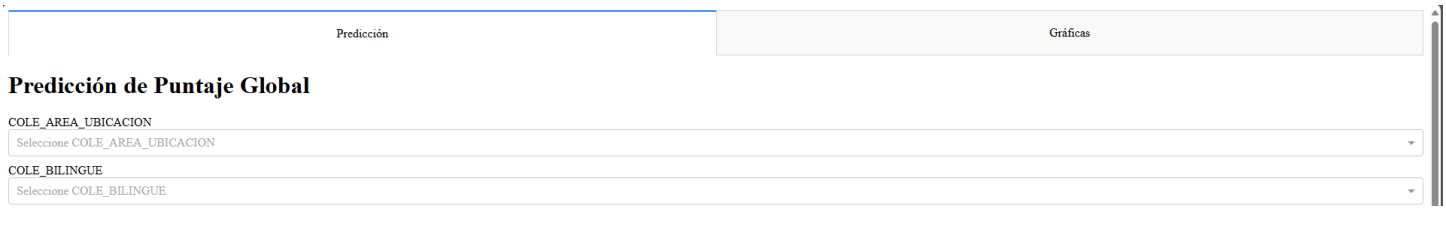
Como se puede ver, los valores de validación se ajustan a los de entrenamiento, lo cual indica que el modelo aprende correctamente de las características proporcionadas y no incurre en sobreajuste.

Ciencia de datos: **Sebastián Arango Crispín.**

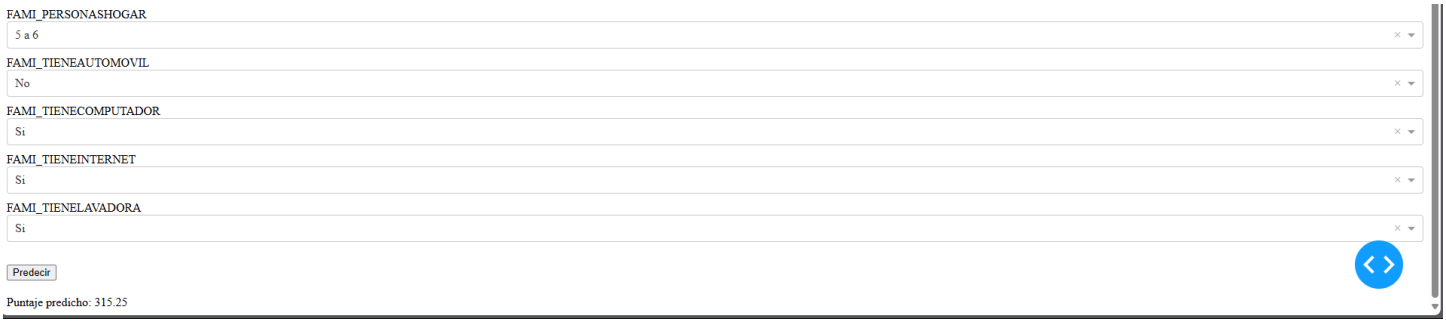
4. Producto

4.1 Diseño y desarrollo del tablero

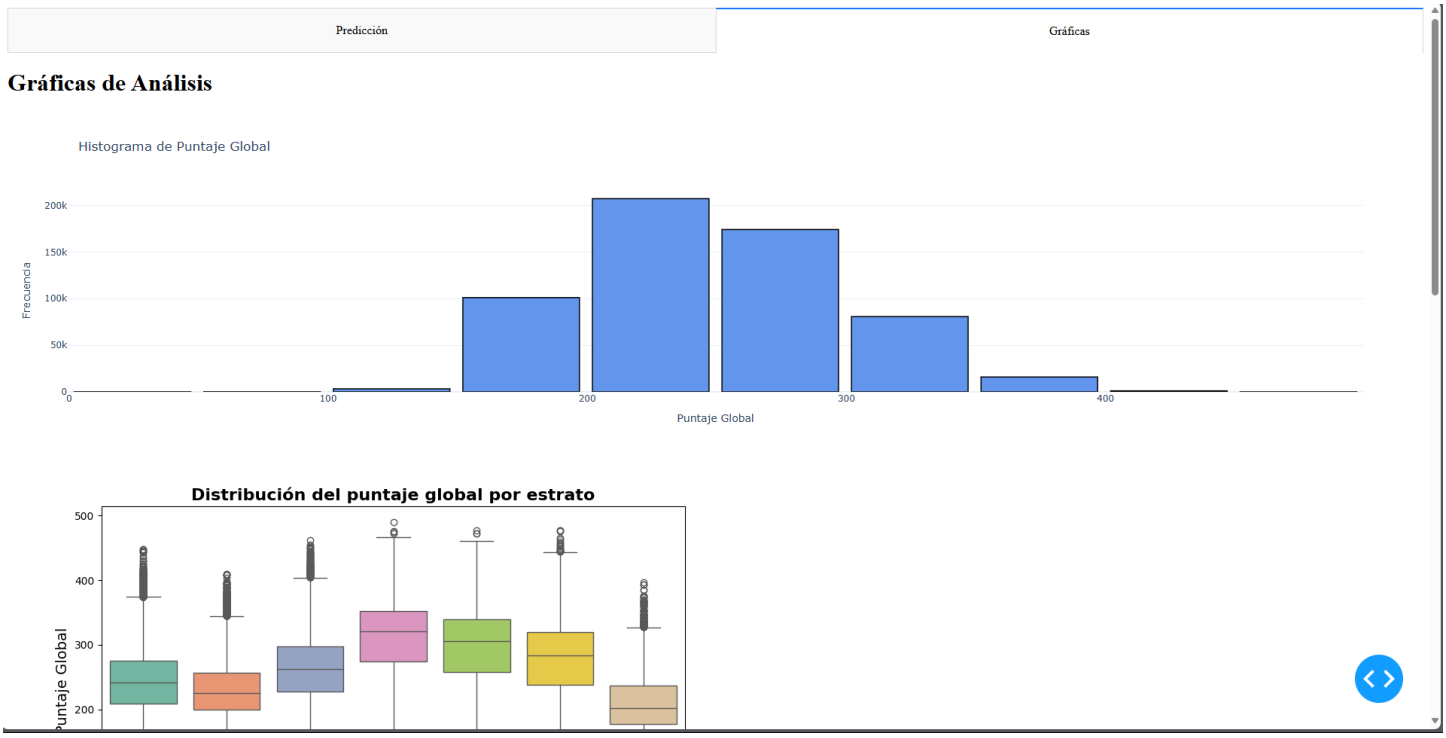
Se diseñó un tablero que tuviera dos ventanas emergentes: una con el modelo de predicción y la otra con una selección de gráficas de los datos como se ve en la siguiente figura:



Para la parte del modelo de predicción se presenta una interfaz de usuario en la cual se ingresan las variables de entrada con las cuales se entrenó el modelo y te muestra la predicción del puntaje global de la prueba como se muestra a continuación:



Por último se muestra una serie de gráficas informativas de los datos iniciales del ejercicio en las cuales se puede ver gráficamente qué factores tienen un mayor efecto en los resultados globales de la prueba.



Diseñador del tablero de datos: **Tomás Liévano Casas.**

4.2 Despliegue

Para realizar el despliegue, se configuró una instancia EC2 en AWS, asignándole una IP elástica para facilitar la conexión desde el dispositivo local usando una clave de acceso. Se optó por una instancia Linux t2.medium, ya que el Dash necesitaba descargar varias bibliotecas que consumen más memoria y requieren un mayor rendimiento. También se subieron los archivos esenciales a la instancia para asegurar que el Dash pudiera establecer la conexión con el modelo y ejecutar las predicciones de manera adecuada.

Link del dash: http://127.0.0.1:8050/

Encargado del despliegue: **Sofía Buitrago Carvajal.**

Repositorio GitHub: https://github.com/arangseb/Proyecto-3-ACTA.git