

Despliegue de soluciones analíticas – MIAD

Proyecto – Entrega final – Grupo 11

Nombre del proyecto

Análisis de los resultados de Evaluar para Avanzar 2022-1 para los estudiantes de instituciones educativas de los departamentos de la Amazonía colombiana.

1. Contexto y problema

1.1. Contexto

En Colombia, el Instituto Colombiano para la Evaluación de la Educación Icfes, es la entidad encargada de evaluar a nivel nacional las habilidades y competencias de los estudiantes en los distintos niveles escolares, desde grado 3° hasta la formación profesional. Para cumplir con este propósito, el ICFES diseña y aplica un conjunto de exámenes estandarizados conocidos como Pruebas Saber (Saber 3°, 5°, 7° y 9°, Saber 11°, Saber TyT y Saber Pro). En los últimos años, el Icfes ha desarrollado una estrategia complementaria a las Pruebas Saber conocida como Evaluar para Avanzar. La implementación de Evaluar para Avanzar se realiza entre los grados 3° y 11°. En estos grados se aplican instrumentos de valoración asociados a las áreas básicas de conocimiento (ver Anexo *Instrumentos que componen Evaluar para Avanzar*).

1.2. Problema

En este sentido, el problema considerado consiste en analizar los resultados de los distintos instrumentos de los estudiantes de las instituciones educativas de los departamentos que componen la Amazonía colombiana (Guainía, Vaupés, Vichada, Guaviare y Amazonas) que aplicaron la estrategia de Evaluar para Avanzar en el 2022-1 e intentar predecir los resultados de la prueba de Lectura a partir de una caracterización del estudiante, considerando que en algunas instituciones no siempre hay disponibilidad de tiempo y recursos para aplicar todos los instrumentos a todos los estudiantes.

2. Pregunta de negocio y alcance del proyecto

La pregunta de negocio busca desarrollar una herramienta analítica que presente los resultados del instrumento de Lectura aplicado durante el 2022-1 de los estudiantes de las instituciones educativas de los departamentos de Amazonas, Guainía, Guaviare, Vaupés y Vichada, así como un modelo predictivo sobre el resultado de este instrumento a partir de características del estudiante, de su establecimiento educativo y los resultados en los demás instrumentos. Se busca así responder a la pregunta:

¿Cuál sería el puntaje en Lectura de un estudiante de alguno de estos 5 departamentos a partir de variables como el género, el grado, la jornada, el municipio y los resultados obtenidos en los otros instrumentos?

2.1. Objetivo del proyecto

El presente proyecto tiene como objetivo analizar los resultados de los estudiantes en la prueba de Lectura de las instituciones educativas colombianas que participaron en la estrategia de Evaluar para Avanzar durante el 2022-1 y utilizarlos junto con otras variables para intentar predecir el resultado en el instrumento de Lectura de estudiantes nuevos o estudiantes que no lograron presentarlo.

2.2. Alcance del proyecto

1. Conocer los resultados de los estudiantes de los establecimientos educativos de 5 departamentos que son parte de la Amazonía colombiana que aplicaron la estrategia de Evaluar para Avanzar durante el 2022-1.
2. Desarrollar un modelo predictivo para el resultado del instrumento de Lectura a partir de variables relacionadas con la naturaleza del establecimiento educativo y con el estudiante y sus resultados en los otros instrumentos.

2.3. Resultados esperados

Como resultado principal del proyecto, se busca obtener un tablero de control que permita visualizar los resultados de los estudiantes de las instituciones educativas, así como una predicción sobre el resultado de este instrumento de Lectura a partir de variables asociadas al estudiante y su establecimiento educativo.

3. Conjunto de datos

Los datos fueron obtenidos a través de la página Web desarrollada por el Instituto Colombiano para la Evaluación de la educación Icfes llamada Datalcfes (<https://www.icfes.gov.co/data-icfes>) que contiene “bases de datos de las distintas pruebas aplicadas para generar conocimiento orientado al mejoramiento y transformación de la calidad de la educación”. De esta página se descargaron los resultados del instrumento conocido como *Evaluar para Avanzar* correspondientes a la aplicación 2022-1. Estos resultados se encuentran en archivos .csv que están organizados en distintas carpetas asociadas a distintas secretarías de educación de ciudades y municipios que aplicaron este instrumento. Para cada observación se cuenta con los siguientes atributos:

- **Estudiante:** Fecha de presentación, Grado y Género
- **Colegio:** Id, Nombre, Sede, Naturaleza, Calendario, Jornada, Municipio, Departamento
- **Resultados:** Modalidad, Instrumento, Componente, Competencia, Cantidad de preguntas total, omitidas e incorrectas.

4. Repositorios en Git y en DVC

4.1. Repositorio en Git

El repositorio del proyecto se encuentra publicado en GitHub y puede ser accedido a través del siguiente enlace: <https://github.com/FabrizioMorales01/DatalcfesAnalysisDSA>.

La estructura de directorios que refleja las diferentes etapas en la metodología CRISP-ML (Cross-Industry Standard Process for Machine Learning) se encuentra definida con detalle en el [repositorio](#):

4.2. Repositorio en DVC

Para el versionamiento de los datos a lo largo del proyecto se ha optado por emplear la herramienta DVC, configurando como respaldo remoto un directorio compartido de Google Drive. Esta configuración presenta varias ventajas para el desarrollo del proyecto (ver Anexo *Repositorio DVC*).

5. Exploración de los datos

5.1. Transformaciones a los datos

En el [diccionario de datos](#) del proyecto se encuentran descritos los atributos con los que se cuenta, junto con su tipo de dato, los cuales podrían requerir algunas transformaciones (Ver anexo *Transformaciones a los datos*). Para 277 registros de estudiantes de los 17.885 que tiene la base de datos (1,55%) se encontró que la cantidad de preguntas del cuadernillo de Lectura es 40, 60, 80 o 140, lo que no es compatible con el diseño de este instrumento, pues los cuadernillos contienen 20 preguntas. Por esta razón, se tomó la decisión de eliminar estos registros ya que no existe confiabilidad sobre estos resultados. La base resultante contiene 17.608 registros. Adicionalmente, de los 17.608 registros se eliminaron 70 (0,4%) para los cuales la cantidad de respuestas correctas en alguno de los 4 instrumentos considerados (CN, CC, LC y MT) es mayor que 20, debido a que los cuadernillos tienen 20 preguntas.

5.2. Análisis descriptivo de los datos

Al analizar la distribución por departamento, encontramos los siguientes registros:

Departamento	Registros	Porcentaje de participación
Guaviare	5.570	31,76 %
Vichada	4.124	23,51 %
Amazonas	3.351	19,11 %
Guainía	3.140	17,90 %
Vaupés	1.353	7,72 %
Total	17.538	100 %

Distribución por Géneros: Si analizamos la distribución por géneros, podemos notar que hay una cantidad similar de estudiantes de ambos géneros: 9.004 mujeres (F) y 8.980 hombres (M).

Departamento	Mujeres	Hombres
Guaviare	2.865	2.705
Vichada	2.080	2.044
Amazonas	1.658	1.693
Guainía	1.547	1.593
Vaupés	640	713
Total	8.790	8.748

Distribución por Grados: Al analizar la distribución por grados en cada departamento, notamos que hay más registros para los grados 3°, 4°, 5° y 6° y que a partir de estos grados empieza a disminuir el número de registros, siendo la menor cantidad la de grado 11°. (Ver anexo *Exploración de los datos figura 4*).

Distribución de Respuestas Correctas de Lectura: Al analizar la distribución del número de respuestas correctas encontramos que la mayoría de los estudiantes tiene un número de respuestas correctas entre 5 y 11. La distribución por departamento es algo sesgada hacia la izquierda, lo que indica que hay menos estudiantes con un número alto de respuestas correctas.

Al analizar la ubicación de los cuartiles, notamos que el departamento del Guaviare presenta un mejor desempeño (leve) y el departamento de Guainía un desempeño levemente menor que los demás. Al analizar la distribución por género, observamos que los departamentos de Guaviare y Vichada presentan un desempeño levemente mayor para las mujeres que para los hombres, en particular en la parte baja de la cantidad de respuestas correctas. Para los otros tres departamentos (Guainía, Vaupés y Amazonas) el número de respuestas correctas de hombres y mujeres es similar. (Ver anexo *Exploración de los datos figura 5*).

Distribución de Respuestas Correctas de Ciencias Naturales: Al analizar la distribución del número de respuestas correctas encontramos que la mayoría de los estudiantes tiene un número de respuestas correctas entre 4 y 9. Se observa que muy pocos estudiantes superan las 10 respuestas correctas. Al analizar la ubicación de los cuartiles, notamos que el departamento del Guaviare presenta un mejor desempeño y el departamento de Vaupés un desempeño levemente menor que los demás. (Ver anexo *Exploración de los datos figura 6*).

Distribución de Respuestas Correctas de Competencias Ciudadanas: Al analizar la distribución del número de respuestas correctas encontramos que la mayoría de los estudiantes tiene un número de respuestas correctas entre 4 y 9. Se observa que muy pocos estudiantes superan las 9 respuestas correctas. Al analizar la distribución por género, observamos que los departamentos de Guaviare y Vichada presentan un desempeño levemente mayor para las mujeres que para los hombres, en particular en la parte alta de la cantidad de respuestas correctas. Para los otros dos departamentos (Vaupés y Amazonas) el número de respuestas correctas de hombres es similar al de las mujeres. (Ver anexo *Exploración de los datos figura 7*).

Distribución de Respuestas Correctas de Matemáticas: Al analizar la distribución del número de respuestas correctas encontramos que la mayoría de los estudiantes tiene un número de respuestas correctas entre 3 y 9. Se observa que muy pocos estudiantes superan las 10 respuestas correctas. Al analizar la ubicación de los cuartiles, notamos que los departamentos del Guaviare y Vaupés presentan un leve mejor desempeño. Al analizar la distribución por género, observamos que los departamentos de Guainía y Vaupés presentan un desempeño levemente mayor para las mujeres que para los hombres, en particular en la parte baja de la cantidad de respuestas correctas. Para los otros dos departamentos (Amazonas, Guaviare y Vichada) el número de respuestas correctas de hombres es mayor que el de las mujeres. (Ver anexo *Exploración de los datos figura 8*).

Distribución de respuestas correctas por jornada: Finalmente, analizamos la distribución del número de respuestas correctas de cada instrumento:

- Para Lectura notamos comportamientos similares entre las jornadas MAÑANA y ÚNICA, un leve desempeño menor para la jornada COMPLETA y un leve desempeño mayor para la jornada TARDE. (Ver anexo *Exploración de los datos figura 9*).

- Para Ciencias Naturales notamos comportamientos similares entre las jornadas MAÑANA y TARDE, un leve desempeño menor para la jornada COMPLETA y un desempeño mayor para la jornada ÚNICA. (Ver anexo *Exploración de los datos figura 10*).
- Para Competencias Ciudadanas notamos comportamientos similares entre las jornadas COMPLETA y TARDE y un desempeño mayor para las jornadas ÚNICA y MAÑANA. (Ver anexo *Exploración de los datos figura 11*).
- Para Matemáticas notamos comportamientos similares entre las jornadas MAÑANA, TARDE y ÚNICA, un leve desempeño mayor para la jornada TARDE. (Ver anexo *Exploración de los datos figura 12*).

6. Modelos de predicción desarrollados

En el informe de la entrega 2 se presentó el trabajo realizado por el equipo para la selección de variables, en el que se definieron 5 escenarios (ver anexo *Selección de variables*). En cuanto a los modelos para la predicción de la cantidad de respuestas correctas del instrumento de Lectura, el equipo consideró los siguientes algoritmos de manera inicial (Todos los modelos considerados fueron implementados utilizando la librería *scikit learn*, junto con el apoyo de las librerías *pandas* y *numpy*):

- Modelo de regresión lineal múltiple
- Modelo de red neuronal
- Modelo de bosque aleatorio (*Random Forest*)
- Modelo de árbol de decisión (*Decision Tree*)
- Modelo de soporte vectorial (*Support Vector Machines*)

6.1. Métricas de desempeño

Para evaluar el desempeño de los modelos considerados bajo los distintos algoritmos y escenarios se eligieron dos métricas: El Score y el MSE. A partir de estas métricas, se identificó que los modelos de bosque aleatorio (*Random Forest*) eran los de mejor desempeño entre las opciones consideradas. Por esta razón, el equipo decidió centrarse en estos algoritmos para refinar los modelos buscando mejor poder predictivo evitando el sobreajuste observado en experimentos previos.

6.2. Modelos finales

A partir de los escenarios planteados y la elección del algoritmo de bosque aleatorio, el equipo implementó dos escenarios y tres modelos para desarrollar, evaluar y elegir el más adecuado para incorporar en la solución analítica desplegada en el tablero:

Escenario 5

Para cada uno de los 4 instrumentos considerados (Ciencias Naturales CN, Competencias Ciudadanas CC, Lectura Crítica LC y Matemáticas MT), se creó una variable que categoriza la cantidad de respuestas correctas de la siguiente manera:

- Desempeño muy bajo (0): Entre 0 y 4 respuestas correctas.
- Desempeño bajo (1): Entre 5 y 9 respuestas correctas.
- Desempeño alto (2): Entre 10 y 14 respuestas correctas.
- Desempeño muy alto (3): Entre 15 y 20 respuestas correctas.

Escenario 6

Para cada uno de los 4 instrumentos considerados (Ciencias Naturales CN, Competencias Ciudadanas CC, Lectura Crítica LC y Matemáticas MT), se creó una variable que categoriza la cantidad de respuestas correctas de la siguiente manera:

- Desempeño bajo (0): Entre 0 y 6 respuestas correctas.
- Desempeño medio (1): Entre 7 y 13 respuestas correctas.
- Desempeño alto (2): Entre 14 y 20 respuestas correctas.

Sobre los algoritmos a implementar, el equipo decidió desarrollar tres basados en árboles de decisión debido a su capacidad para reconocer patrones complejos. A partir de las categorías definidas en los escenarios anteriores, los algoritmos implementados fueron de clasificación:

- **Random Forest:** algoritmo de ensamble que combina un conjunto de árboles de decisión independientes. Cada árbol de decisión se entrena en un subconjunto aleatorio de los datos de entrenamiento.
- **Gradient Boosting:** algoritmo de ensamble que combina un conjunto de árboles de decisión en serie. Cada árbol de decisión se entrena para reducir el error del modelo anterior. Esto ayuda a mejorar la precisión del modelo.
- **XGBoost:** variante de *Gradient Boosting* que utiliza una serie de optimizaciones para mejorar la velocidad y la precisión del modelo.

6.3. Evaluación de los modelos

Una vez definidos los escenarios, los algoritmos y las métricas de desempeño en cuestión, cada integrante del equipo procedió a implementar un algoritmo diferente utilizando ML FLOW. Se realizaron más de 48 corridas durante las últimas semanas, mientras se refinaba, tanto el código como los modelos. Finalmente, nos concentramos en 12 corridas, correspondientes a 6 escenarios con cada algoritmo (Ver anexo *Experimentos en MLFlow*). A continuación, se presentan los resultados de los experimentos realizados.

Escenario 5	Score	MSE
Random Forest	0,535	0,635
Gradient Boosting	0,572	0,532
XGBoost	0,581	0,522

Escenario 6	Score	MSE
Random Forest	0,603	0,440
Gradient Boosting	0,635	0,383
XGBoost	0,639	0,385

En ambos escenarios, se observa que el algoritmo de XGBoost presenta un mejor desempeño en ambas métricas. Por esta razón, se decidió aplicar *GridSearchCV* para encontrar los parámetros óptimos, que fueron los siguientes:

Optimización de parámetros	n_estimators	learning_rate	max_depth
Escenario 5	200	0.2	5
Escenario 6	200	0.2	4

A partir de los resultados obtenidos, el equipo seleccionó el modelo XGBoost bajo el escenario 6 para incorporar en la solución analítica que será desplegada en el tablero de control.

6.4. Modelos complementarios

Además del desarrollo de modelos para la predicción del desempeño en Lectura Crítica, el equipo también implementó modelos predictivos para el desempeño en Matemáticas, en Ciencias Naturales y en Competencias Ciudadanas, bajo los mismos dos escenarios planteados previamente (E5 y E6). Al evaluar el desempeño de estos modelos bajo las dos métricas definidas anteriormente, se obtuvieron los siguientes resultados:

Matemáticas	Score_E5	MSE_E5	Score_E6	MSE_E6
Random Forest	0.596	0.488	0.636	0.375
Gradient Boosting	0.620	0.429	0.659	0.353
XGBoost	0.631	0.426	0.657	0.354

Ciencias Naturales	Score_E5	MSE_E5	Score_E6	MSE_E6
Random Forest	0.660	0.473	0.720	0.312
Gradient Boosting	0.682	0.401	0.760	0.263
XGBoost	0.692	0.418	0.750	0.275

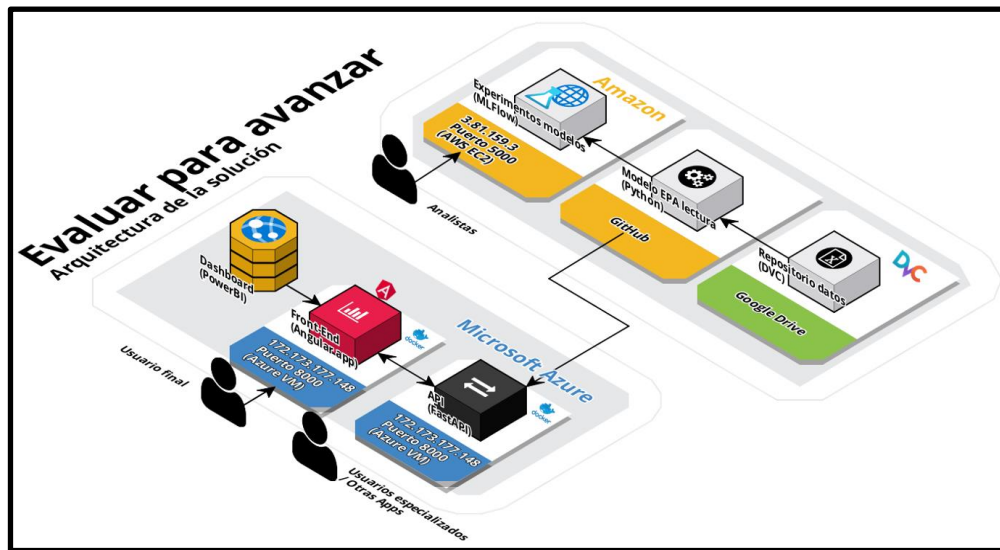
Competencias Ciudadanas	Score_E5	MSE_E5	Score_E6	MSE_E6
Random Forest	0.650	0.485	0.711	0.322
Gradient Boosting	0.670	0.435	0.737	0.284
XGBoost	0.689	0.404	0.746	0.277

Los experimentos se realizaron y registraron usando MLFlow a través de una máquina en AWS EC2, la cual se encuentra en hibernación en caso de que se requiera su reactivación. Los resultados de los experimentos desarrollados se pueden ver en el anexo *Experimentos en MLFlow*.

7. Tablero de control desarrollado

7.1. Despliegue del tablero

Para el despliegue de la solución se optó por la estrategia de contenedores Docker sobre Infraestructura como Servicio ofrecida por Azure, de esta manera se evitaron pasos adicionales en el despliegue que podrían ser propensos a errores y se garantiza incluir en cada contenedor los requerimientos necesarios para que tanto la aplicación FrontEnd como el API se ejecuten sin problema. Esta decisión de arquitectura de la solución también estuvo alineada con las capacidades y conocimientos de los miembros del equipo y la necesidad de mantener estos dos componentes disponibles para los usuarios finales el mayor tiempo posible.



También se consideró como apoyo para la fase de entrenamiento y evaluación de modelos un servicio de IaaS en Amazon, particularmente el servicio EC2 para publicar un servidor MLFlow a través del cual los miembros del equipo pudieran registrar experimentos de los diversos modelos. La razón para tomar esta decisión en la arquitectura de la solución correspondía a un ahorro en costos y a que la necesidad de este servicio era esporádica y no se requería mantener la máquina activa 100% del tiempo, sin embargo, se tuvo en cuenta una configuración especial para permitir la hibernación de la máquina y de esta manera siguieran estando disponibles los datos de los experimentos tras cada inicialización (ver anexo *Tablero de control*).

El tablero se puede acceder a través de: <http://172.173.177.148:8001/pages/dashboard-icfes>. El API a través de http://172.173.177.148:8000/docs/#/default/predict_api_v1_predict_post

7.2. Funcionalidades del tablero

El propósito de esta herramienta es analizar el impacto de Evaluar para Avanzar en las instituciones educativas de la Amazonía Colombiana. Para ello, la herramienta recopila datos de las instituciones educativas que aplican los instrumentos de Evaluar para Avanzar en los departamentos de Guaviare, Guainía, Vaupés, Vichada y Amazonas, y realiza un análisis de

estos datos para identificar tendencias y patrones. Estos datos incluyen información sobre la institución, los estudiantes evaluados, y los resultados de los instrumentos de competencias ciudadanas, ciencias naturales, matemáticas y nuestro foco: lectura crítica. El análisis puede incluir dimensiones como región, naturaleza, calendario o jornada de la institución; género. Finalmente, presenta los resultados del análisis de forma visual para facilitar su comprensión. La herramienta incluye gráficos y tablas que permiten visualizar los datos de forma clara y concisa.

En la parte superior se muestra un diagrama de cajas con estadísticos como cuartiles 1 y 3, media, mediana, máximo y mínimo de respuestas correctas por departamento. Al costado derecho resumen de cantidad de colegios, estudiantes evaluados y los promedios separados de cada uno de los instrumentos en medición. Allí se encuentran opciones de segmentación por nombre del colegio, municipio, jornada e instrumento con las que se puede configurar la visualización del resto del panel. En el centro del tablero, la cantidad de estudiantes evaluados por cada grado y departamento y, la distribución de estudiantes por género. En el costado inferior, un gráfico de cintas que entrega información del promedio de respuestas correctas que, además, las jerarquiza de mayor a menor permitiendo identificar el departamento con mejores y peores resultados en cada uno de los grados. Finalmente, un gráfico de distribución geoespacial con información de cantidad de colegios de naturaleza oficial o no oficial, en cada departamento. (Ver anexo *Tablero de control*).

En la esquina inferior derecha, se encuentra el botón para acceder a los comandos necesarios para realizar la predicción, en la ventana flotante se ingresa la información conocida del estudiante al que se quiere realizar la proyección, grado y género del estudiante; nombre, modalidad y jornada del colegio donde estudia y las respuestas correctas que obtuvo en las pruebas de ciencias naturales, competencias ciudadanas y matemáticas, se presiona el botón verde “Predecir” y se obtiene el nivel esperado del estudiante en los resultados de Lectura Crítica que puede ser bajo, medio o alto. (Ver anexo *Tablero de control*).

8. Resultados y conclusiones

En la realización de este proyecto, se probaron varios conjuntos de variables y algoritmos para desarrollar modelos de predicción para el desempeño de los estudiantes de Lectura Crítica. Finalmente, se implementó un modelo utilizando el algoritmo XGBoost, el cual demostró un mejor desempeño comparado con los algoritmos de Random Forest y Gradient Boosting. Dentro de las variables consideradas para el modelo se incluyeron el género, grado y jornada escolar y el desempeño en los otros tres instrumentos.

Además, se creó un panel interactivo que ofrece información relevante sobre los resultados de los estudiantes de las instituciones educativas en los departamentos de la Amazonía colombiana. Este panel se convierte en una herramienta práctica para tomar decisiones informadas y generar hipótesis sobre el impacto del programa Evaluar para Avanzar en la región, contribuyendo a una comprensión más profunda de los factores que afectan su rendimiento en las distintas áreas evaluadas. El panel de control creado resulta ser una herramienta útil tanto para las autoridades educativas como para las partes involucradas, ya



que ofrece una interfaz adecuada para analizar y comprender los resultados de Evaluar para Avanzar en la Amazonía colombiana contribuyendo a la toma de decisiones informadas en la política educativa.



Por último, en cuanto a la posibilidad de diseñar y desplegar nuevos tableros o crear nuevos espacios en el tablero existente, se tiene la opción de incluir modelos para predicciones no solo del instrumento de Lectura Crítica sino de los instrumentos de Matemáticas, de Ciencias Naturales y de Competencias Ciudadanas, siendo estas 2 últimas las que tienen modelos con mejor capacidad de predicción. Así mismo, se puede considerar la incorporación del análisis para los 32 departamentos de Colombia, incluso para municipios o Entidades Territoriales Certificadas en particular.

9. Reporte del trabajo en equipo

Estructura del equipo y responsabilidades

El equipo está conformado por 4 estudiantes de la Maestría en Inteligencia Analítica de Datos, del curso Despliegue de soluciones Analíticas. La organización del equipo es la siguiente:

 **Project Administrator:** Celimo Fabricio Morales
 **Principal Investigator:** Daniel Alfonso Londoño

 **Data Manager:** Dayron Alberto Cuadros
 **Contributor:** David Mauricio Ruiz

Dinámica de trabajo

El equipo decidió organizar 2 reuniones semanales (martes y viernes entre las 6 p.m. y las 10 p.m.) con el fin de monitorear el avance de las actividades establecidas en el proyecto, así como acordar el trabajo que cada integrante debe desarrollar semanalmente y resolver las dudas, inconvenientes e inquietudes que hayan surgido durante las actividades. (Ver Anexos > Progreso)

Es posible acceder a la programación de las actividades por cada iteración semanal a través del siguiente enlace: <https://github.com/users/dayroncj/projects/4/views/4>

10. Enlace al video:

El video puede ser consultado a través del enlace:

https://drive.google.com/file/d/1O3hi7F7_o2YJI6jwOopzVoUR78xNuzF/view?usp=sharing

11. Bibliografía

ICFES, 2020. Infografía General de Evaluar para Avanzar. Bogotá D.C. Último acceso: sábado 28 de 2023. <https://www.icfes.gov.co/web/guest/hist%C3%B3rico-2020>

Anexos

Instrumentos que componen Evaluar para Avanzar

La siguiente figura muestra la distribución de los instrumentos a lo largo de los grados.

Instrumento de valoración	Grados									
	3º	4º	5º	6º	7º	8º	9º	10º	11º	
Competencias Comunicativas en Lenguaje: Lectura										
Matemáticas										
Ciencias Naturales y Educación Ambiental										
Competencias Ciudadanas: Pensamiento Ciudadano										
Inglés										
Cuestionarios Auxiliares										

Figura 1. Distribución de los instrumentos de valoración por grados.

Fuente: Infografía General de Evaluar para Avanzar, ICFES, 2020.

Transformaciones a los datos

Omitir o remover atributos correlacionados de los resultados de las pruebas	<p>Atributos a omitir/remover:</p> <ul style="list-style-type: none"> EXA_N_PREGUNTAS, número de preguntas en total para el instrumento de Lectura. EXA_N_PREGUNTAS_OM, número de preguntas que el/la estudiante se abstuvo de responder. EXA_PRC_PREGUNTAS_OM, porcentaje de preguntas que el/la estudiante se abstuvo de responder. EXA_N_RTAS_CORR, número de respuestas correctas en el instrumento de Lectura. EXA_N_RTAS_NOCORR, número de respuestas incorrectas en el instrumento de Lectura. EXA_PRC_RTAS_NOCORR, porcentaje de respuestas no correctas en el instrumento de Lectura. EXA_CUADERNILLO, cuadernillo aplicado. EXA_COMPONENTE, componente asociado al instrumento. EXA_INSTRUMENTO, área de conocimiento valorada por el instrumento de valoración, en este caso Lectura. EXA_COMPETENCIA, competencia evaluada con el instrumento, en este caso Comprensión lectora.
Aplicar one hot encoding a atributos categóricos	<ul style="list-style-type: none"> ESTU_GRADO, grado que cursa el estudiante ESTU_GENERO, género del estudiante COLE_NATURALEZA, naturaleza del establecimiento educativo COLE_CALENDARIO, calendario de la sede COLE_JORNADA, Jornada de la sede EXA_MODALIDAD, modalidad de presentación del instrumento de valoración
Identificación de nuevos atributos	<p>A partir de la base de datos original, se crearon las siguientes variables:</p> <ul style="list-style-type: none"> EXA_N_RTAS_CORR_CN, cantidad de respuestas correctas en el instrumento de Ciencias Naturales EXA_N_RTAS_CORR_CC, cantidad de respuestas correctas en el instrumento de Competencias Ciudadanas

	<ul style="list-style-type: none"> • EXA_N_RTAS_CORR_LC, cantidad de respuestas correctas en el instrumento de Lectura Crítica • EXA_N_RTAS_CORR_MT, cantidad de respuestas correctas en el instrumento de Matemáticas
--	--

Tabla 1. Transformaciones a los datos.

Exploración de los datos

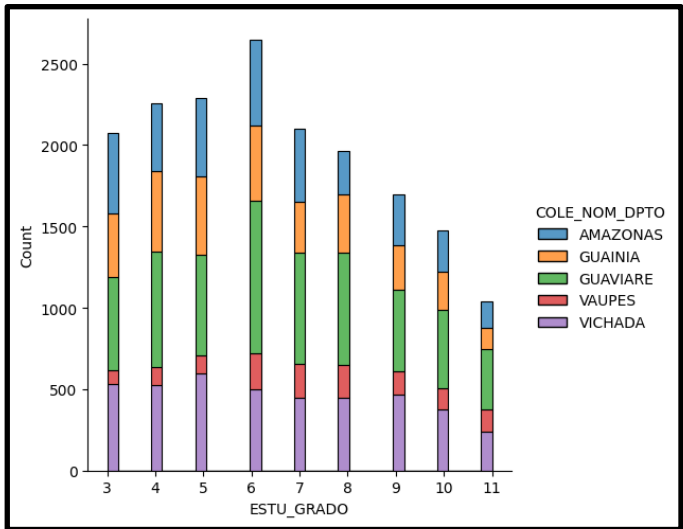


Figura 4. Distribución de registros por grado y departamento.
Fuente: Elaboración propia

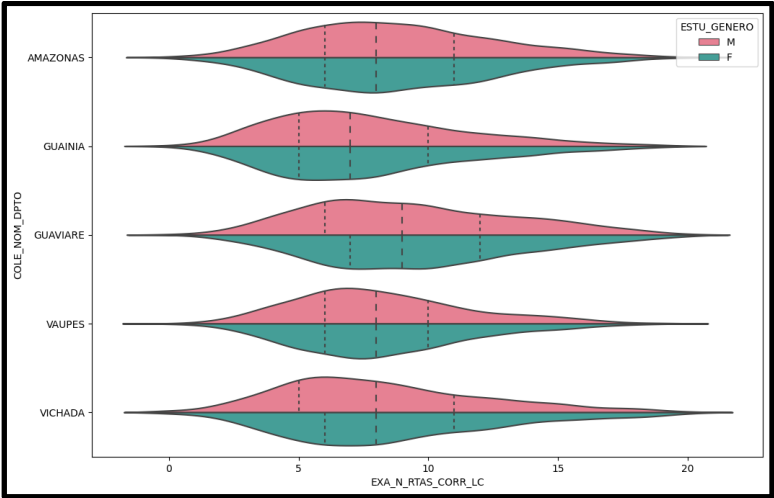


Figura 5. Distribución del número de respuestas correctas de Lectura por departamento y género.
Fuente: Elaboración propia

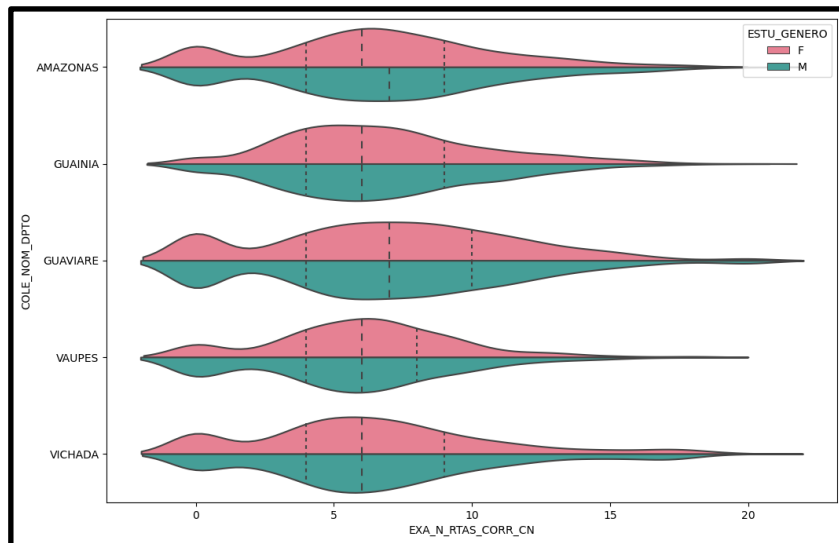


Figura 6. Distribución del número de respuestas correctas de Ciencias Naturales por departamento y género. Fuente: Elaboración propia

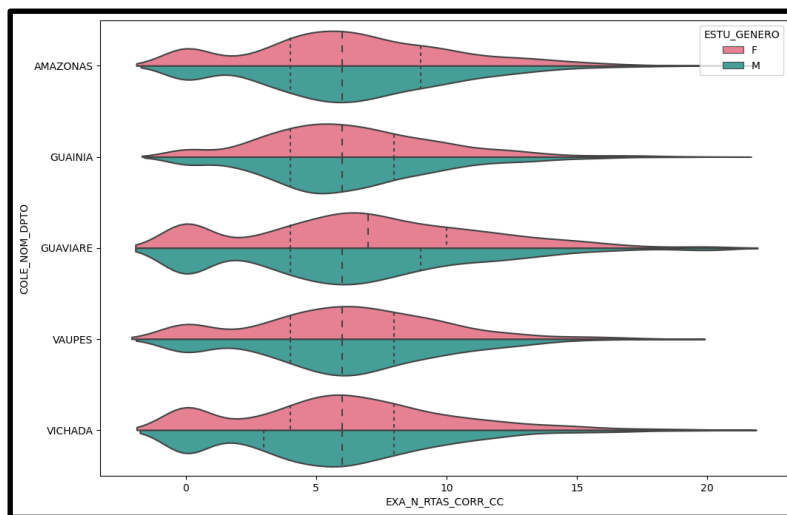


Figura 7. Distribución del número de respuestas correctas de Competencias ciudadanas por departamento y género. Fuente: Elaboración propia

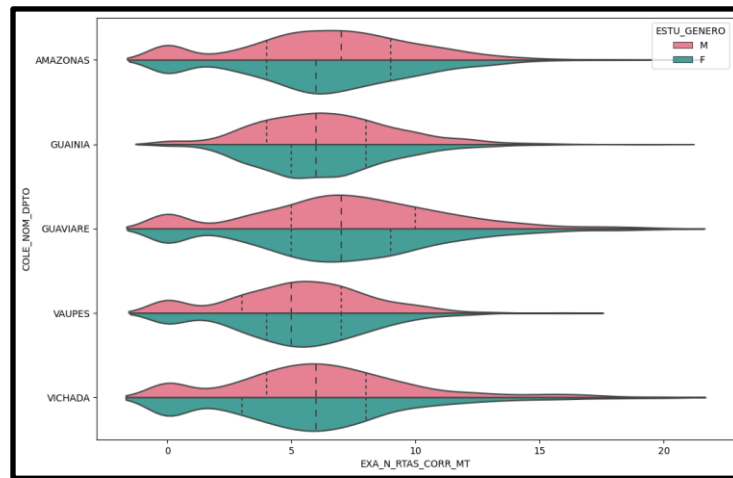


Figura 8. Distribución del número de respuestas correctas de Matemáticas por departamento y género.
Fuente: Elaboración propia

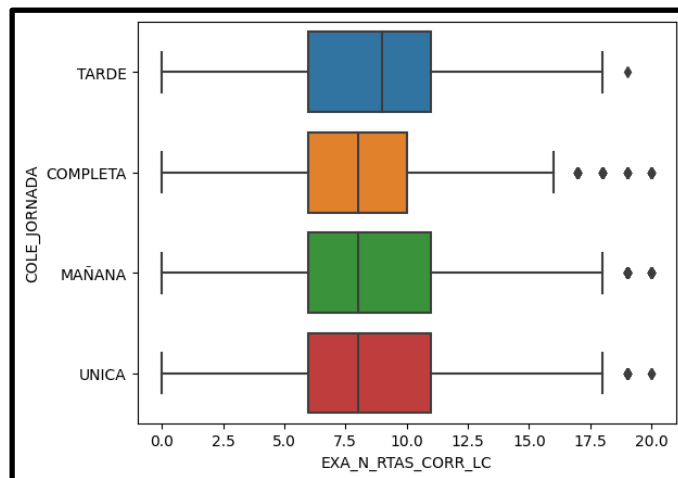


Figura 9. Distribución del número de respuestas correctas de Lectura por jornada.
Fuente: Elaboración propia

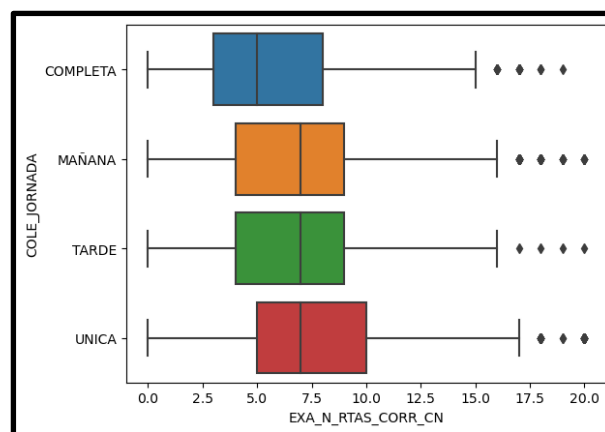


Figura 10. Distribución del número de respuestas correctas de Ciencias Naturales por jornada.
Fuente: Elaboración propia

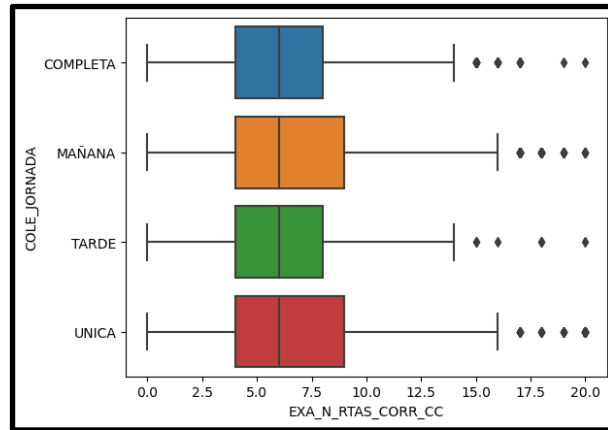


Figura 11. Distribución del número de respuestas correctas de Competencias ciudadanas por jornada.
Fuente: Elaboración propia

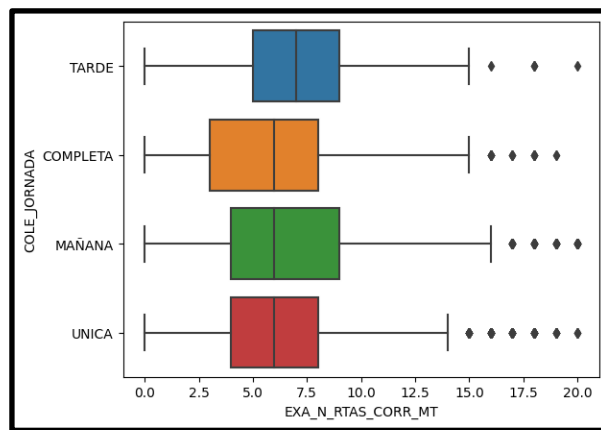


Figura 12. Distribución del número de respuestas correctas de Matemáticas por jornada.
Fuente: Elaboración propia

Repositorio DVC

Las ventajas de usar DVC con Google Drive como repositorio de los datos son:

- Los datos están disponibles en la nube y son accesibles desde cualquier lugar con conexión a internet.
- Google Drive proporciona opciones de seguridad y control de acceso, lo que significa que es posible limitar quién puede ver y editar los datos.
- Google Drive ofrece opciones para aumentar el espacio de almacenamiento según sea necesario.
- DVC y Google Drive facilitan la distribución de datos a través de múltiples entornos de desarrollo o servidores de procesamiento.

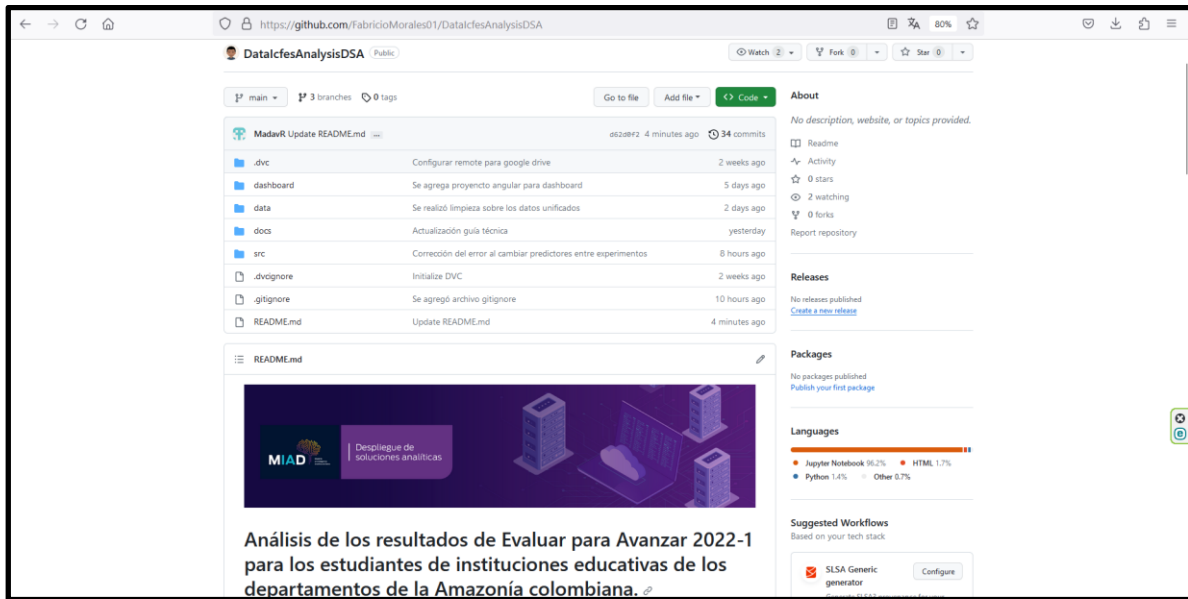


Figura 3. Repositorio en GitHub del proyecto.
Fuente: Elaboración propia

Selección de variables

Para la selección de variables, el equipo consideró importante analizar el poder predictivo de las variables numéricas asociadas a los instrumentos que presentó el estudiante e identificó la variable asociada al establecimiento educativo como una variable que podría ser muy relevante para la predicción. Desde esta perspectiva, se definieron los siguientes 5 escenarios:

1. Escenario 1 (E1): En este escenario el equipo consideró todas las variables numéricas asociadas al estudiante, al establecimiento educativo, al municipio y al departamento.

- ESTU_GRADO: Corresponde al grado del estudiante (entre 3 y 11).
- COLE_COD_ICFES: Corresponde al código del establecimiento educativo.
- COLE_COD_MUNICIPIO: Corresponde al código del municipio en el que se ubica el establecimiento educativo al que pertenece el estudiante.
- COLE_COD_DEPARTAMENTO: Corresponde al código del departamento en el que se ubica el establecimiento educativo al que pertenece el estudiante.
- EXA_N_RTAS_CORR_CN: Corresponde al número de respuestas correctas del estudiante en el instrumento de Ciencias Naturales.
- EXA_N_RTAS_CORR_CC: Corresponde al número de respuestas correctas del estudiante en el instrumento de Competencias Ciudadanas.
- EXA_N_RTAS_CORR_MT: Corresponde al número de respuestas correctas del estudiante en el instrumento de Matemáticas.

2. Escenario 2 (E2): En este escenario el equipo consideró solamente las variables numéricas asociadas directamente al estudiante sin incluir variables asociadas al establecimiento educativo, al municipio ni al departamento:

- ESTU_GRADO: Corresponde al grado del estudiante (entre 3 y 11).
- EXA_N_RTAS_CORR_CN: Corresponde al número de respuestas correctas del estudiante en el instrumento de Ciencias Naturales.
- EXA_N_RTAS_CORR_CC: Corresponde al número de respuestas correctas del estudiante en el instrumento de Competencias Ciudadanas.
- EXA_N_RTAS_CORR_MT: Corresponde al número de respuestas correctas del estudiante en el instrumento de Matemáticas.
- COLE_COD_ICFES: Corresponde al código del establecimiento educativo.

3. Escenario 3 (E3): En este escenario el equipo consideró las siguientes variables numéricas, omitiendo la variable asociada al colegio para analizar la capacidad predictora al considerar solamente el departamento y el municipio asociado al estudiante:

- ESTU_GRADO: Corresponde al grado del estudiante (entre 3 y 11).
- COLE_COD_MCPIO: Corresponde al código del municipio en el que se ubica el establecimiento educativo al que pertenece el estudiante.
- COLE_COD_DEPTO: Corresponde al código del departamento en el que se ubica el establecimiento educativo al que pertenece el estudiante.
- EXA_N_RTAS_CORR_CN: Corresponde al número de respuestas correctas del estudiante en el instrumento de Ciencias Naturales (entre 0 y 20).
- EXA_N_RTAS_CORR_CC: Corresponde al número de respuestas correctas del estudiante en el instrumento de Competencias Ciudadanas (entre 0 y 20).
- EXA_N_RTAS_CORR_MT: Corresponde al número de respuestas correctas del estudiante en el instrumento de Matemáticas (entre 0 y 20).

4. Escenario 4 (E4): En este escenario el equipo consideró todas las variables , tanto numéricas como categóricas, asociadas al estudiante, el establecimiento educativo, el municipio y el departamento:

- ESTU_GRADO: Corresponde al grado del estudiante (entre 3 y 11).
- COLE_COD_ICFES: Corresponde al código del establecimiento educativo.
- COLE_COD_MCPIO: Corresponde al código del municipio en el que se ubica el establecimiento educativo al que pertenece el estudiante.
- COLE_COD_DEPTO: Corresponde al código del departamento en el que se ubica el establecimiento educativo al que pertenece el estudiante.
- EXA_N_RTAS_CORR_CN: Corresponde al número de respuestas correctas del estudiante en el instrumento de Ciencias Naturales (entre 0 y 20).
- EXA_N_RTAS_CORR_CC: Corresponde al número de respuestas correctas del estudiante en el instrumento de Competencias Ciudadanas (entre 0 y 20).
- EXA_N_RTAS_CORR_MT: Corresponde al número de respuestas correctas del estudiante en el instrumento de Matemáticas (entre 0 y 20).
- ESTU_GENERO: Indica el estudiante es de género femenino o masculino
- COLE_NATURALEZA: Indica si el establecimiento educativo es OFICIAL o NO OFICIAL.
- COLE_CALEDARIO: Indica si el establecimiento educativo es de calendario A, calendario B u OTRO.
- COLE_JORNADA: Indica la jornada del establecimiento educativo (COMPLETA, MAÑANA, NOCHE o SABATINA).
- EXA_MODALIDAD: Indica la modalidad de presentación de los instrumentos (ONLINE, OFFLINE, CUADERNILLOS Y PDF, PAPEL Y LÁPIZ).

5. Escenario 5 (E5): En este escenario el equipo consideró todas las variables , tanto numéricas como categóricas, asociadas al estudiante, el establecimiento educativo, el municipio y el departamento. Sin embargo, a diferencia del escenario 4 se estandarizaron las variables:

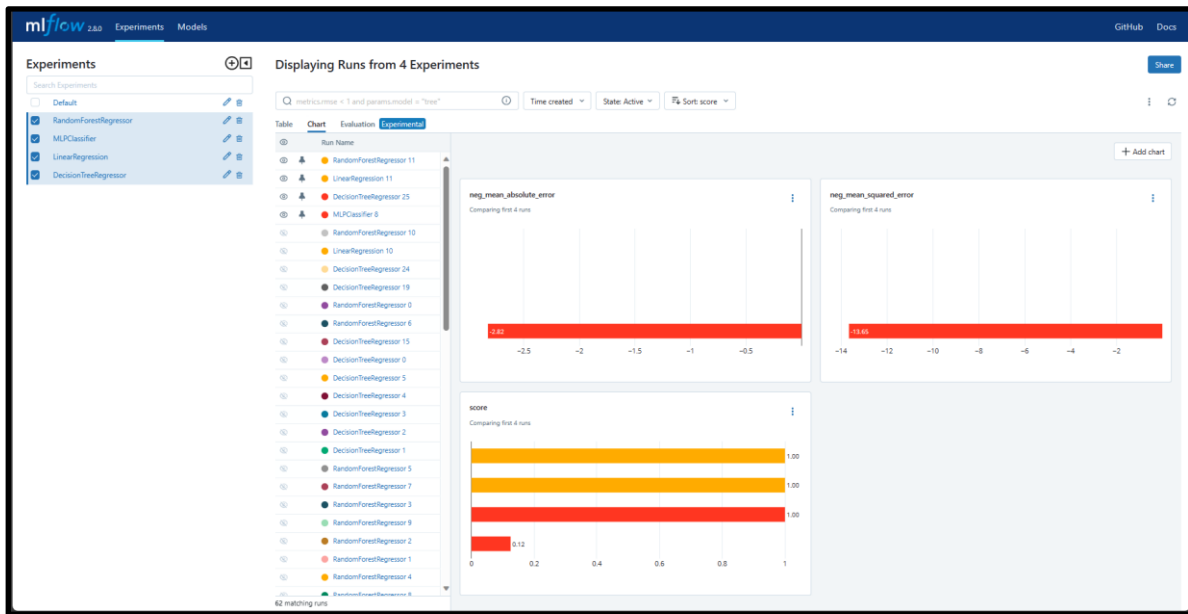
- ESTU_GRADO: Corresponde al grado del estudiante (entre 3 y 11).
- COLE_COD_ICFES: Corresponde al código del establecimiento educativo.
- COLE_COD_MCPIO: Corresponde al código del municipio en el que se ubica el establecimiento educativo al que pertenece el estudiante.
- COLE_COD_DEPTO: Corresponde al código del departamento en el que se ubica el establecimiento educativo al que pertenece el estudiante.
- EXA_N_RTAS_CORR_CN: Corresponde al número de respuestas correctas del estudiante en el instrumento de Ciencias Naturales (entre 0 y 20).
- EXA_N_RTAS_CORR_CC: Corresponde al número de respuestas correctas del estudiante en el instrumento de Competencias Ciudadanas (entre 0 y 20).
- EXA_N_RTAS_CORR_MT: Corresponde al número de respuestas correctas del estudiante en el instrumento de Matemáticas (entre 0 y 20).
- ESTU_GENERO: Indica el estudiante es de género femenino o masculino
- COLE_NATURALEZA: Indica si el establecimiento educativo es OFICIAL o NO OFICIAL.
- COLE_CALENDARIO: Indica si el establecimiento educativo es de calendario A, calendario B u OTRO.
- COLE_JORNADA: Indica la jornada del establecimiento educativo (COMPLETA, MAÑANA, NOCHE o SABATINA).
- EXA_MODALIDAD: Indica la modalidad de presentación de los instrumentos (ONLINE, OFFLINE, CUADERNILLOS Y PDF, PAPEL Y LÁPIZ).

Experimentos en MLFlow

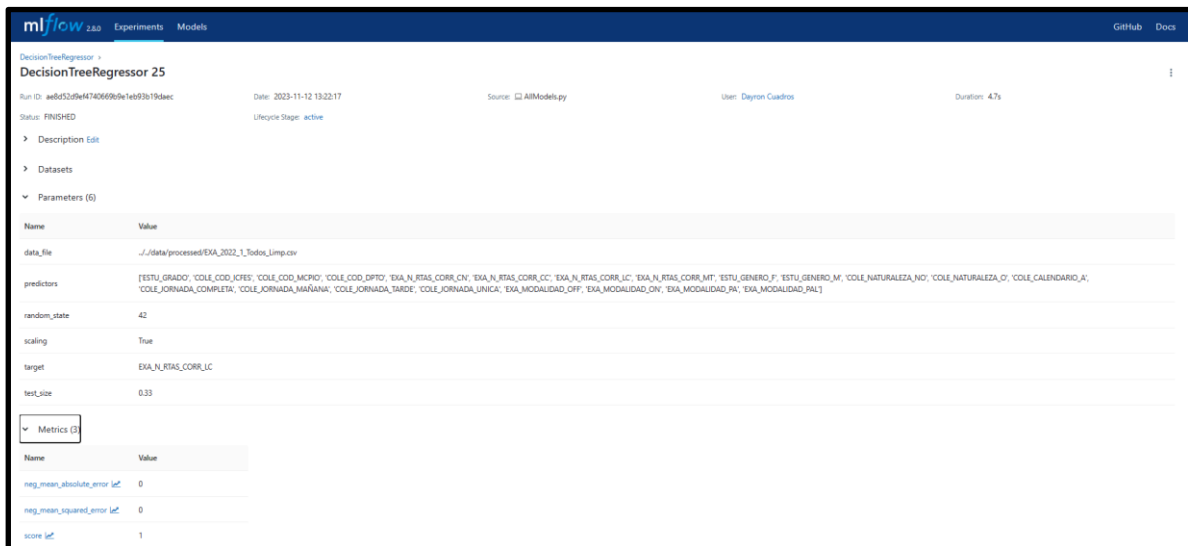
Displaying Runs from 4 Experiments

Run Name	Created	Duration	neg_mean_absolute_error	neg_mean_squared_error	score	activation	data_file	hidden_layer_size	max_depth	max_iter	n_estimators	predictions
RandomForestRegressor 11	38 seconds ago	8.0s	0	0	1	-	./data/pro...	-	80	-	200	[ESTI...
LinearRegression 11	2 minutes ago	5.1s	-1.25340458...	-2.42272285...	1	-	./data/pro...	-	-	-	-	[ESTI...
DecisionTreeRegressor 25	3 minutes ago	4.7s	0	0	1	-	./data/pro...	-	-	-	-	[ESTI...
MLPClassifier 8	2 minutes ago	9.8s	-2.82360055...	-13.8559936...	0.12491361...	relu	./data/pro...	(10, 10)	-	1000	-	[ESTI...
RandomForestRegressor 10	46 seconds ago	8.2s	0	0	1	-	./data/pro...	-	80	-	200	[ESTI...
LinearRegression 10	2 minutes ago	4.9s	-1.25340458...	-2.42272285...	1	-	./data/pro...	-	-	-	-	[ESTI...
DecisionTreeRegressor 24	3 minutes ago	4.3s	0	0	1	-	./data/pro...	-	-	-	-	[ESTI...
DecisionTreeRegressor 19	17 minutes ago	10.5s	0	0	1	-	./data/pro...	-	-	-	-	[ESTI...
RandomForestRegressor 9	1 hour ago	22.2s	-2.44526728...	-8.99379041...	0.13386044...	-	./data/pro...	-	80	-	200	[ESTI...
RandomForestRegressor 6	1 minute ago	18.2s	-2.44674793...	-10.0030895...	0.1340737...	-	./data/pro...	-	80	-	200	[ESTI...
DecisionTreeRegressor 15	36 minutes ago	11.8s	-2.55910815...	-10.4518795...	0.13240964...	-	./data/pro...	-	80	-	-	[ESTI...
DecisionTreeRegressor 9	1 hour ago	5.7s	-2.55910815...	-10.4518795...	0.13240964...	-	./data/pro...	-	80	-	-	[ESTI...
DecisionTreeRegressor 5	1 hour ago	5.6s	-2.57169796...	-10.5261773...	0.27730857...	-	./data/pro...	-	80	-	-	[ESTI...
DecisionTreeRegressor 4	1 hour ago	5.3s	-2.57169796...	-10.5261773...	0.27730857...	-	./data/pro...	-	80	-	-	[ESTI...
DecisionTreeRegressor 3	1 hour ago	5.8s	-2.57169796...	-10.5261773...	0.27730857...	-	./data/pro...	-	80	-	-	[ESTI...
DecisionTreeRegressor 2	1 hour ago	5.2s	-2.57169796...	-10.5261773...	0.27730857...	-	./data/pro...	-	80	-	-	[ESTI...
DecisionTreeRegressor 1	1 hour ago	4.8s	-2.57169796...	-10.5261773...	0.27730857...	-	./data/pro...	-	80	-	-	[ESTI...
RandomForestRegressor 5	1 hour ago	17.4s	-2.56794874...	-10.7217988...	0.16387787...	-	./data/pro...	-	80	-	200	[ESTI...
RandomForestRegressor 7	1 minute ago	14.8s	-2.56820438...	-10.7399117...	0.16263435...	-	./data/pro...	-	80	-	200	[ESTI...
RandomForestRegressor 3	1 hour ago	17.8s	-2.56719887...	-10.7412863...	0.16263413...	-	./data/pro...	-	80	-	200	[ESTI...
RandomForestRegressor 9	1 minute ago	15.8s	-2.56847742...	-10.7419896...	0.16249307...	-	./data/pro...	-	80	-	200	[ESTI...
RandomForestRegressor 2	1 hour ago	18.4s	-2.56854404...	-10.7515104...	0.16183797...	-	./data/pro...	-	80	-	200	[ESTI...
RandomForestRegressor 1	1 hour ago	18.1s	-2.57006138...	-10.7554148...	0.16156993...	-	./data/pro...	-	80	-	200	[ESTI...

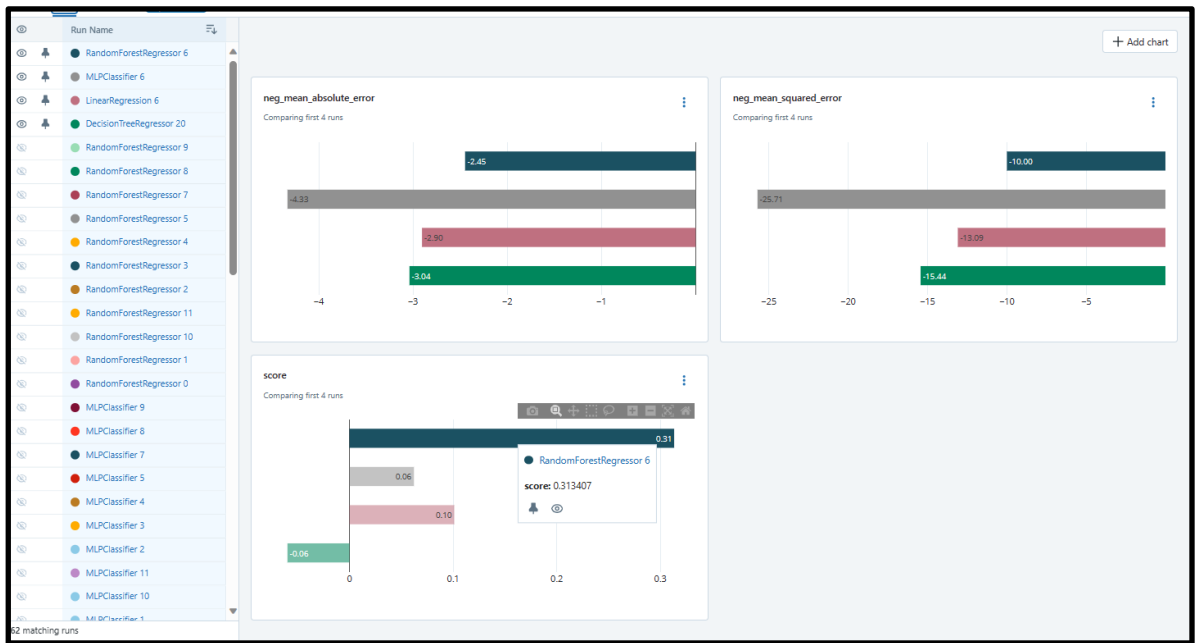
Selección de las mejores corridas por algoritmo ordenadas por la métrica de score



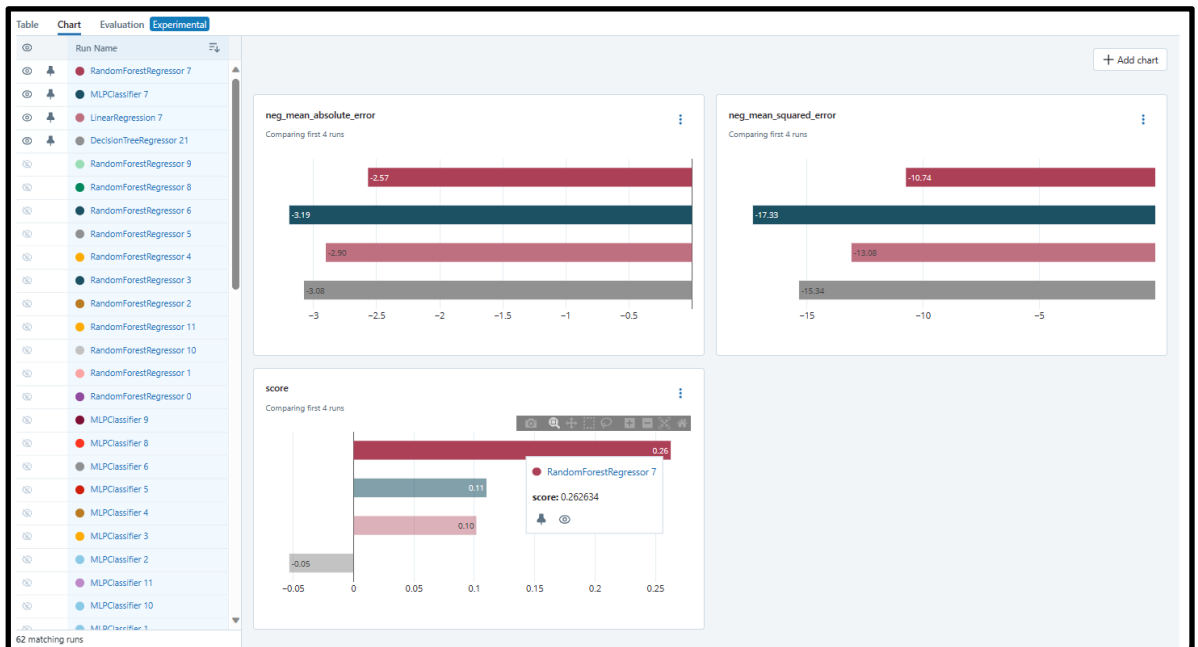
Comparación de métricas (MSE, MAE, Score) de los 4 mejores



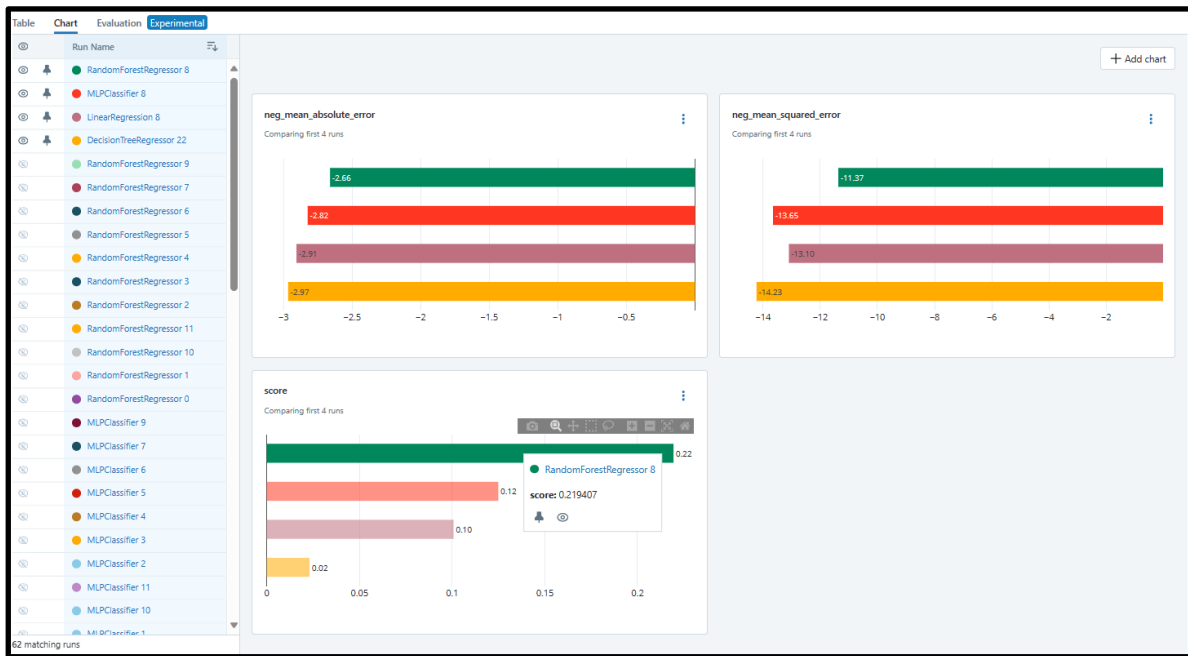
Revisión de los parámetros y métricas particulares de la mejor corrida (RandomForest escenario 5)



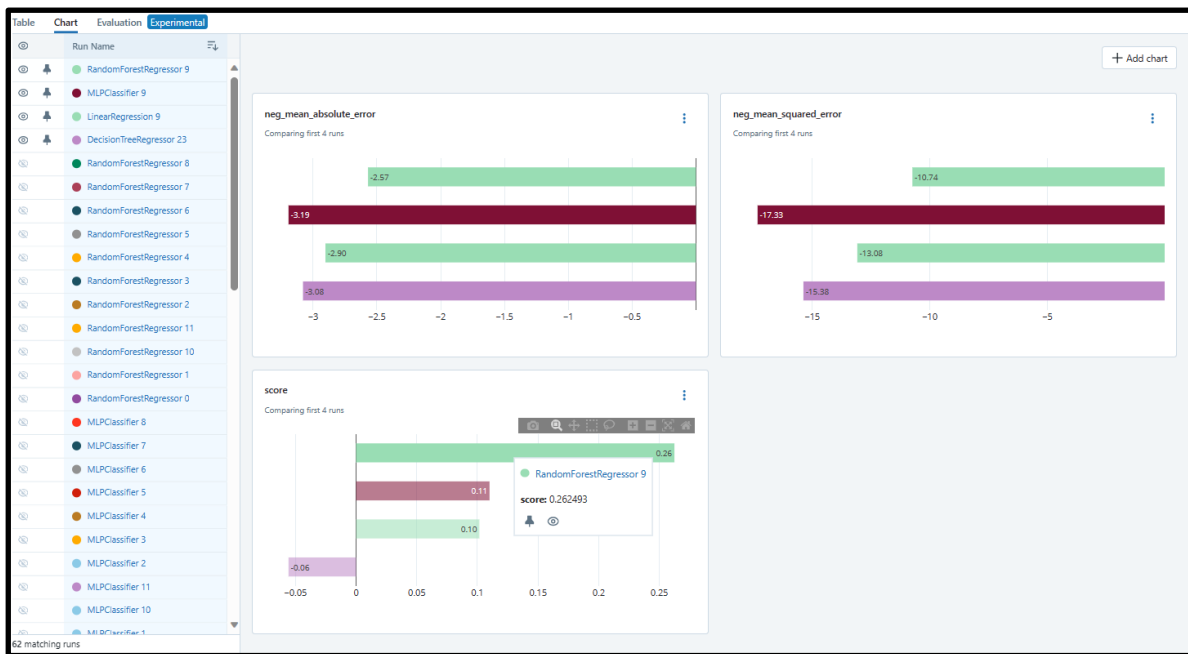
Comparación escenario 0



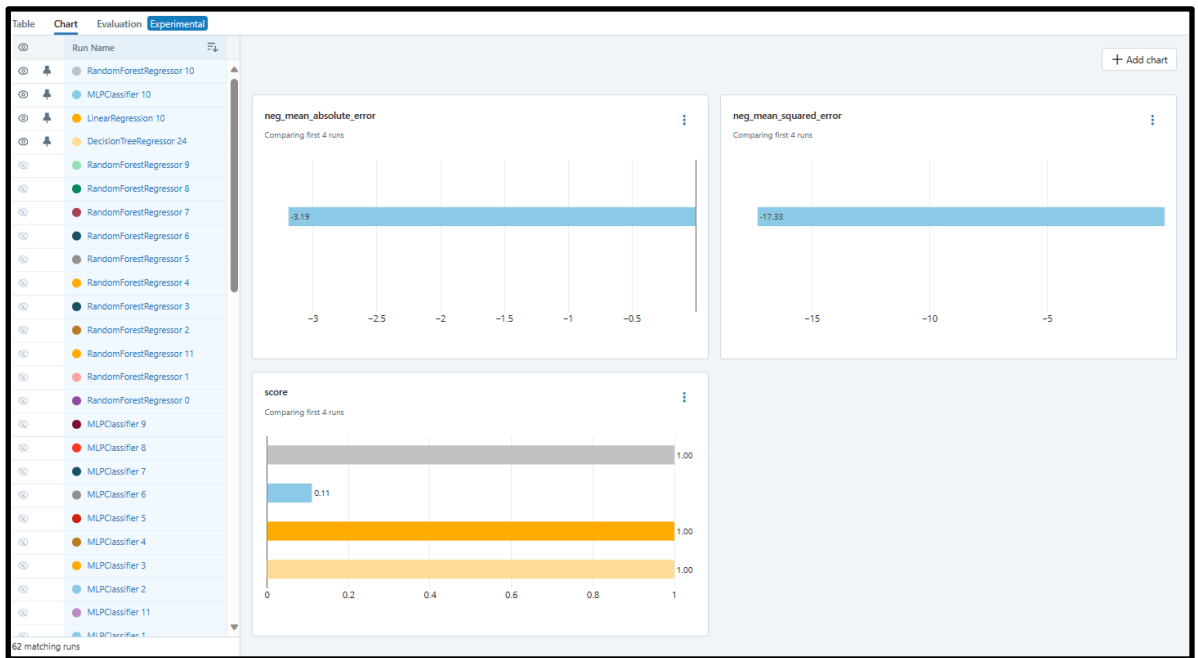
Comparación escenario 1



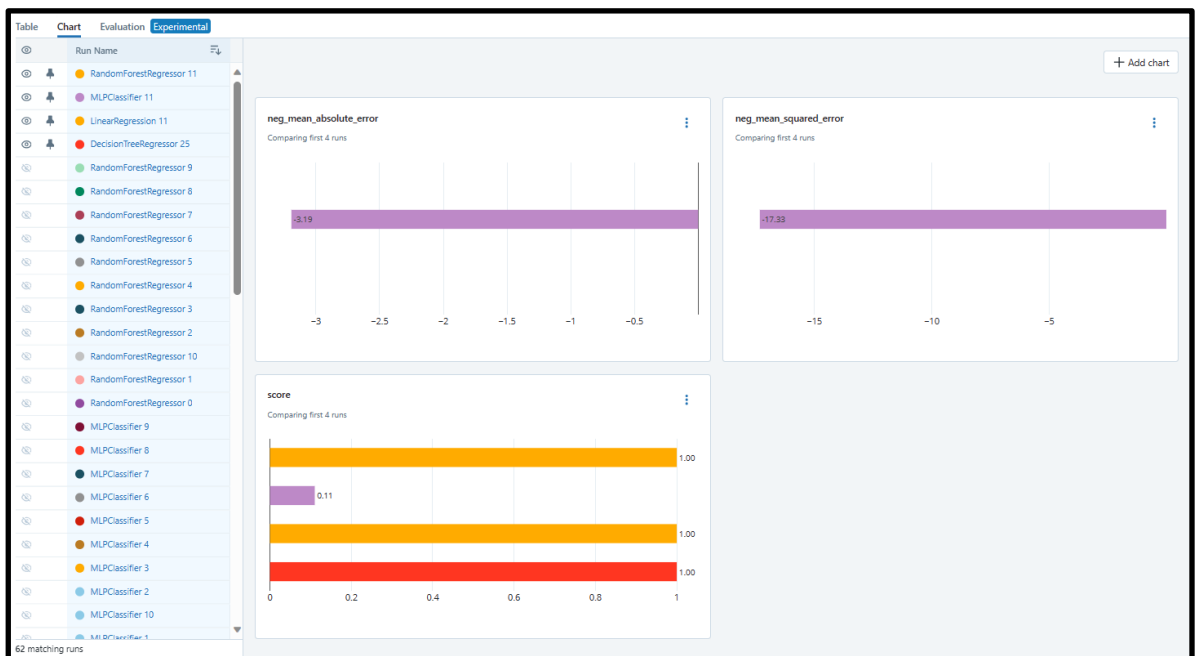
Comparación escenario 2



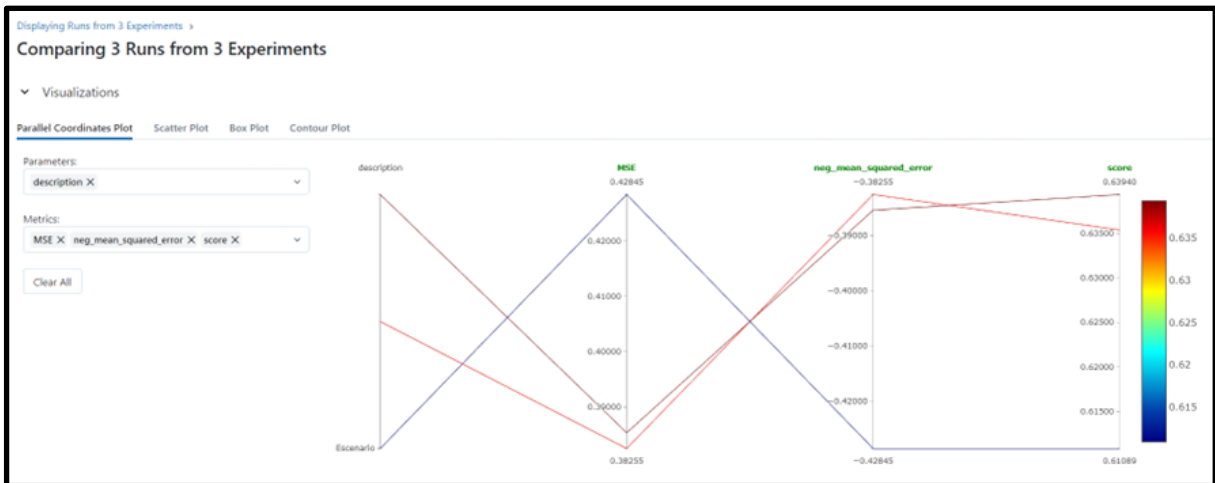
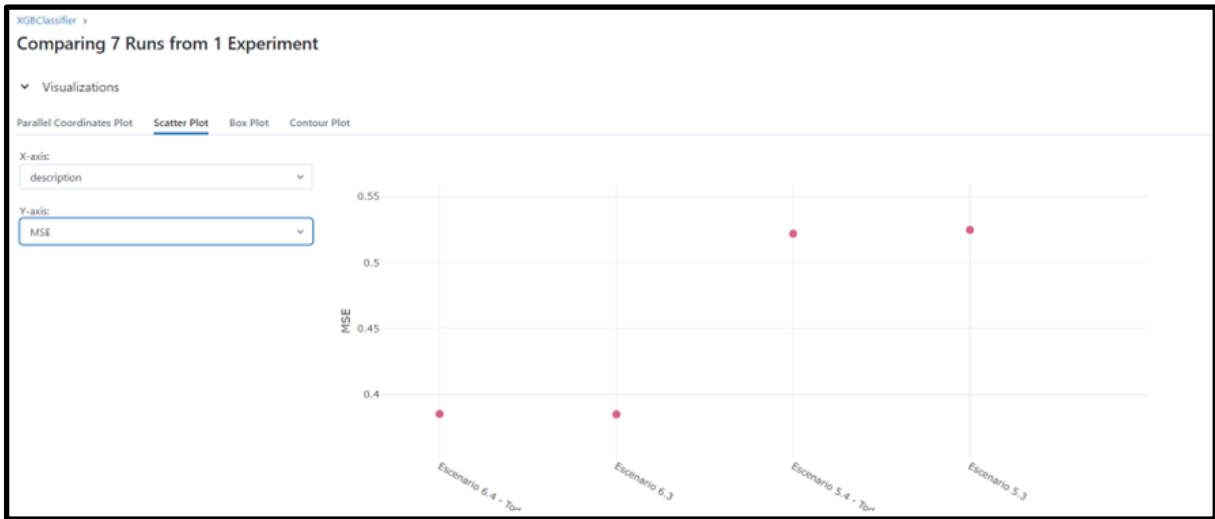
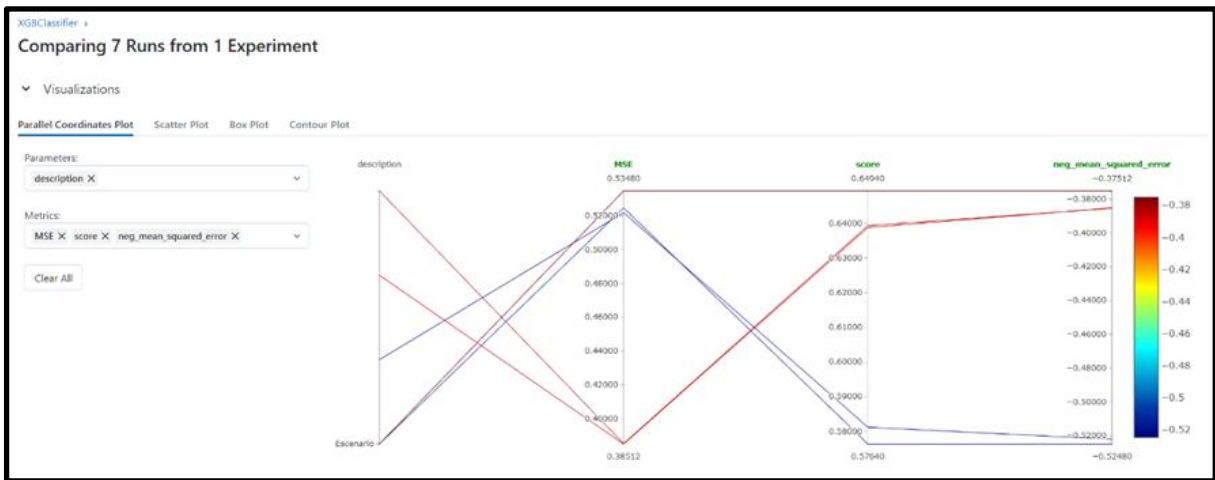
Comparación escenario 3



Comparación escenario 4



Comparación escenario 5



				Metrics				
<input type="checkbox"/>	Run Name	Created	Duration	MSE		neg_mean_absol	neg_mean_squa	score
<input type="checkbox"/>	RandomForestClassifier 5	1 hour ago	14.7s	0.428		-0.402	-0.428	0.611
<input type="checkbox"/>	XGBClassifier 6	1 hour ago	7.9s	0.385		-0.369	-0.385	0.639
<input type="checkbox"/>	GradientBoostingClassifier 4	1 hour ago	11.4s	0.383		-0.371	-0.383	0.635

Experiments

Search Experiments

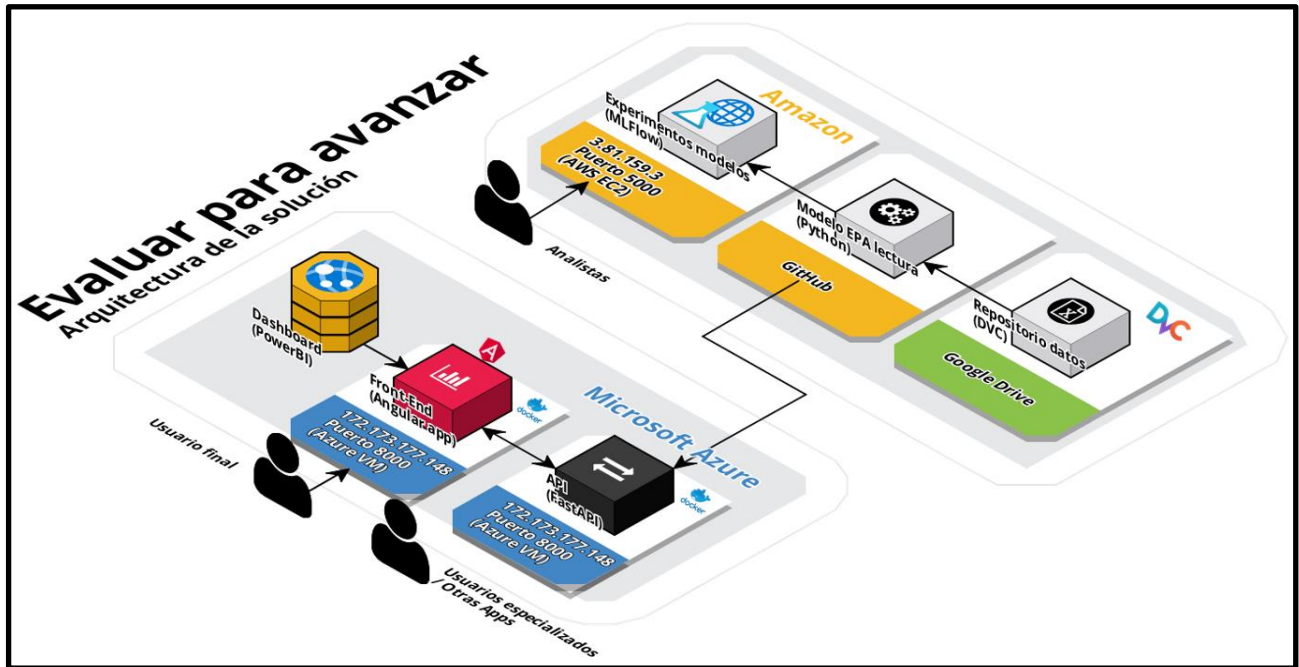
☐ Default
 ☒ DecisionTreeRegressor
 ☒ LinearRegression
 ☒ MLPClassifier
 ☒ RandomForestRegressor
 ☒ RandomForestClassifier
 ☒ GradientBoostingClassifier
 ☒ XGBClassifier

Displaying Runs from 7 Experiments

Run Name		Created	Duration	MSE	neg_mean_absol	neg_mean_squa	score
<input checked="" type="checkbox"/>	GradientBoostingClassifier 4	1 hour ago	11.4s	0.383	-0.371	-0.383	0.635
<input checked="" type="checkbox"/>	XGBClassifier 6	1 hour ago	7.9s	0.385	-0.369	-0.385	0.639
<input checked="" type="checkbox"/>	RandomForestClassifier 5	1 hour ago	14.7s	0.428	-0.402	-0.428	0.611
<input type="checkbox"/>	RandomForestRegressor 11	2 hours ago	10.4s	0	0	0	1
<input type="checkbox"/>	RandomForestRegressor 10	2 hours ago	10.3s	0	0	0	1
<input type="checkbox"/>	RandomForestRegressor 5	2 hours ago	10.7s	0	0	0	1
<input type="checkbox"/>	RandomForestRegressor 4	2 hours ago	10.4s	0	0	0	1
<input type="checkbox"/>	MLPClassifier 5	2 hours ago	17.7s	0	0	0	1
<input type="checkbox"/>	DecisionTreeRegressor 5	2 hours ago	7.2s	0	0	0	1
<input type="checkbox"/>	DecisionTreeRegressor 4	2 hours ago	6.5s	0	0	0	1
<input type="checkbox"/>	LinearRegression 4	2 hours ago	6.4s	2.423e-28	-1.252e-14	-2.423e-28	1
<input type="checkbox"/>	XGBClassifier 5	1 hour ago	7.3s	0.385	-0.369	-0.385	0.639
<input type="checkbox"/>	XGBClassifier 4	1 hour ago	8.2s	0.522	-0.452	-0.522	0.581
<input type="checkbox"/>	XGBClassifier 3	1 hour ago	7.8s	0.525	-0.457	-0.525	0.576
<input type="checkbox"/>	GradientBoostingClassifier 3	1 hour ago	12.8s	0.532	-0.462	-0.532	0.572
<input type="checkbox"/>	GradientBoostingClassifier 2	2 hours ago	13.6s	0.532	-0.462	-0.532	0.572
<input type="checkbox"/>	GradientBoostingClassifier 1	2 hours ago	13.6s	0.532	-0.462	-0.532	0.572
<input type="checkbox"/>	GradientBoostingClassifier 0	2 hours ago	13.3s	0.532	-0.462	-0.532	0.572

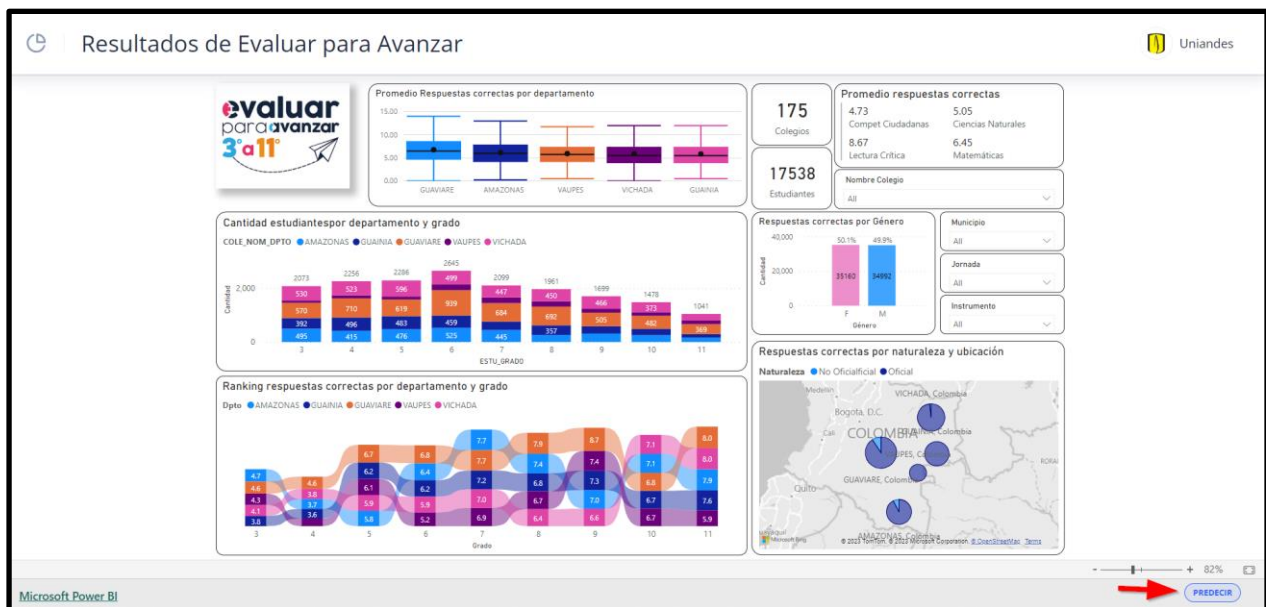
Tablero de control: Despliegue y funcionalidades

Despliegue



Estructura del despliegue del tablero de control

Funcionalidades



Análisis descriptivo

ICFES - Evaluar para avanzar

172.173.177.148:8001/pages/dashboard-icfes

Comenzar a usar Firefox Nueva pestaña Nueva pestaña Nueva pestaña Nueva pestaña

Resultados de Evaluar

Unianides

Predicir respuestas correctas de Lectura

Grado: 4 Género: Femenino

Instituto: C.E. EL TROMPILLO

Modalidad: Online Jornada: Mañana

Respuestas correctas (0-20)

Ciencias Naturales: 1 Competencias Ciudadanas: 0 Matemáticas: 0

Se ha clasificado con un:

Nivel bajo de lectura

PREDECIR

Predicción del nivel bajo de lectura

ICFES - Evaluar para avanzar

172.173.177.148:8001/pages/dashboard-icfes

Comenzar a usar Firefox Nueva pestaña Nueva pestaña Nueva pestaña Nueva pestaña

Resultados de Evaluar

Unianides

Predicir respuestas correctas de Lectura

Grado: 11 Género: Femenino

Instituto: C.E. INDÍGENA CAÑO NEGRO UNO

Modalidad: Offline Jornada: Tarde

Respuestas correctas (0-20)

Ciencias Naturales: 3 Competencias Ciudadanas: 3 Matemáticas: 5

Se ha clasificado con un:

Nivel medio de lectura

PREDECIR

Predicción del nivel medio de lectura

ICFES - Evaluar para avanzar

172.173.177.148:8001/pages/dashboard-icfes

Comenzar a usar Firefox Nueva pestaña Nueva pestaña Nueva pestaña Nueva pestaña

Resultados de Evaluar

Uniaandes

Predicir respuestas correctas de Lectura

Grado: 8 Género: Masculino

Instituto: I.E. COLEGIO TECNICO AGROPECUARIO JOSE CELESTINO MUTIS

Modalidad: Offline Jornada: Tarde

Respuestas correctas (0-20)

Ciencias Naturales: 8 Competencias Ciudadanas: 8 Matemáticas: 8

Se ha clasificado con un:

Nivel alto de lectura

PREDECIR

Microsoft Power BI

Predicción del nivel alto de lectura

Progreso

MIAD DSA Evaluar para avanzar

Home Current iteration Next iteration Planning

Filter by keyword or by field Discard

Title	Status	Assignee	CRISP-DM stage	
Semana 5 Nov 06 - Nov 12 Current				
16 Refinamiento del problema de negocio	Done	David Ruiz	Entendimiento de los datos	
17 Actualización de notebook de EDA y documentación	Done	David Ruiz	Entendimiento de los datos	
18 Preprocesamiento de nuevos datos y acotamiento a 5 departamentos de región Amazonía	Done	David Ruiz, Dayron Cuadros	Preparación de los datos	
19 Limpieza de datos	Done	David Ruiz	Preparación de los datos	
20 Desarrollo de nuevos modelos	Done	David Ruiz	Modelado	
21 Refinamiento de la implementación del dashboard en PowerBI	Done	Daniel Londoño	Modelado	
22 Primera versión de la aplicación Angular	Done	Fabrizio Morales	Despliegue	
23 Primera versión de la API para servir datos de modelos descriptivos y predictivos	Done	Dayron Cuadros	Despliegue	
24 Actualización documentación en repositorio	Done	Team		
25 Documentación segunda entrega	Done	Team		
Semana 6 Nov 13 - Nov 19				
26 Desarrollar nuevas versiones de los modelos	Backlog		Modelado	
27 Comparar y seleccionar mejores alternativas. Emplear Mlflow para versionar los modelos y los re...	Backlog		Modelado	
28 Empaquetar y desplegar una primera versión del tablero y los modelos.	Backlog		Despliegue	
29 Documentación de los avances de la semana 6	Backlog			

Progreso de las Semanas 5 y 6

▼ Semana 7 4 Nov 20 - Nov 26					
31	⌚ Desarrollar nuevas versiones de los modelos	✓ Done	-	David Ruiz	Modelado -
32	⌚ Comparar y seleccionar mejores alternativas. Emplear MLflow para versionar los modelos y los r...	✓ Done	-	Dayron Cuadros	Modelado -
33	⌚ Documentación de los avances de la semana 7	✓ Done	-	Team	-
34	⌚ Continuación del empaquetamiento y despliegue del tablero y los modelos	✓ Done	-	Fabrizio Morales	-
+ Add item					
▼ Semana 8 2 Nov 27 - Dec 03 Current					
35	⌚ Empaquetar, integrar y desplegar la última versión del tablero y los modelos	✓ Done	-	Team	Despliegue -
36	⌚ Actualización de la documentación en repositorio e informe	✓ Done	-	Team	-
+ Add item					

Progreso de las Semanas 7 y 8