**Universidad de los Andes**

Logotipo, Icono

El contenido generado por IA puede ser incorrecto.

**Analítica para la Toma de Decisiones**

**2025-1**

**Profesor:**

**Juan Fernando Pérez**

**Manuela Lizcano (202122826), Juan Diego Lozano (202122869), Felipe Rivera (202123500)**

**Ingeniería Industrial**

**Roles:**

**Juan Diego Lozano: Análisis de negocio y tablero de datos**

**Felipe Rivera M: Ingeniería de datos y análisis de datos**

**Manuela Lizcano: Ciencia de datos, despliegue y mantenimiento**

# Análisis del negocio (Juan Diego)

El objetivo de este proyecto es lograr desarrollar un producto analítico, usando datos sobre las Pruebas Saber 11, específicamente usando los datos del año 2015. A partir de un modelo de clasificación, queremos analizar el desempeño de los estudiantes en las diferentes competencias que son evaluadas en el icfes. Específicamente con el proyecto queremos desarrollar unos tableros didactas con los que podemos responder a las siguientes preguntas de interés:

## Preguntas de negocio:

1. ¿Influye el nivel educativo de los padres en el desempeño del estudiante en las pruebas Saber 11?

Solución descriptiva: Lo que queremos llegar a realizar con esta pregunta es lograr comparar los puntajes de los estudiantes según el nivel educativo de la madre y del padre. De manera más detallada queremos analizar el desempeño de cada estudiante en cada una de las áreas evaluadas. Las cuales específicamente son: Ciencias Naturales, Matemáticas, Lectura Critica, Inglés y Sociales. Tras realizar este análisis queremos identificar si existe una relación entre mejor desempeño de los estudiantes según el perfil académico de sus padres.

1. ¿Qué características del entorno familiar y del colegio están más asociadas con un desempeño que le permita al estudiante ingresar a la Universidad de los Andes?

Solución predictiva: A partir del modelo de clasificación se quieren analizar las variables sociodemográficas de cada uno de los estudiantes, con para identificar si tienen las cualidades de un perfil que pudiese obtener un puntaje para ingresar a la Universidad de los Andes. Este modelo únicamente analizar el ingreso a las siguientes carreras universitarias:

* Administración, Biología, Física, Geociencias, Matemáticas, Química, Economía, Ingeniería Ambiental, Ingeniería Civil, Ingeniería Biomédica, Ingeniería de Sistemas y Computación, Ingeniería Electrónica, Ingeniería Electrónica

Ingeniería Industrial, Ingeniería Mecánica e Ingeniería Química.

El objetivo de esta solución predictiva es lograr generar un tablero interactivo donde los estudiantes recién graduados de colegio lograran analizar su capacidad para el ingreso a la Universidad de los Andes, según su desempeño en las Pruebas Saber.

# Exportar datos desde aws

# Exploración de Datos

Ilustración 1. Correlaciones de las variables seleccionadas

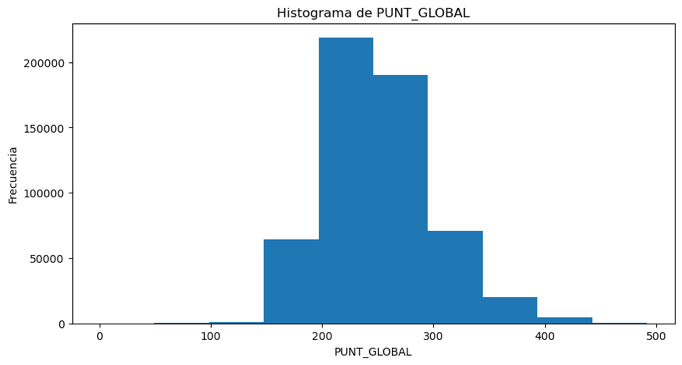


Ilustración 2. Histograma puntaje global

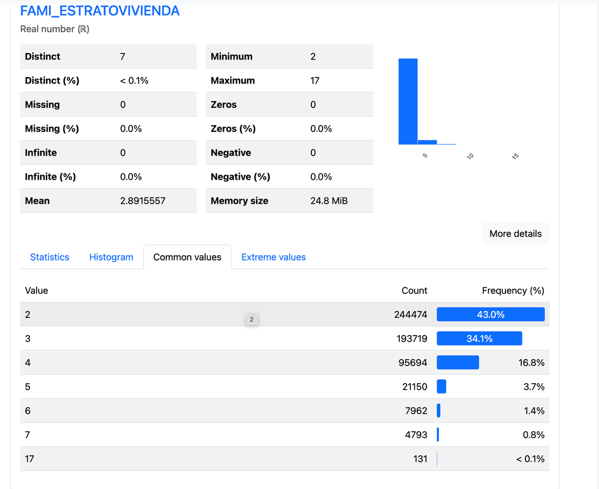


Ilustración 3. Análisis estrato de vivienda

Durante la etapa de exploración de datos se aplicó un análisis estadístico utilizando Pandas Profiling. Este análisis permitió entender la distribución de las variables y sus relaciones más significativas. Como parte de este proceso, se generaron histogramas, gráficos de correlación y resúmenes estadísticos que ayudaron a identificar patrones.

Por ejemplo, al observar la variable FAMI\_ESTRATOVIVIENDA se evidenció una fuerte concentración en los estratos 2 y 3, que representan el 43% y 34.1% respectivamente de los datos, revelando una posible representación sesgada hacia estratos medios. Además, la matriz de correlación mostró relaciones importantes entre variables socioeconómicas como la tenencia de computador o automóvil y el puntaje global del estudiante. Estos hallazgos fueron claves para decidir qué variables incluir en el modelo predictivo.

# Limpieza de Datos

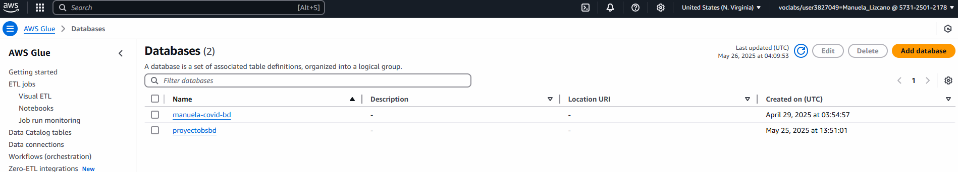


Ilustración 4-AWS\_1

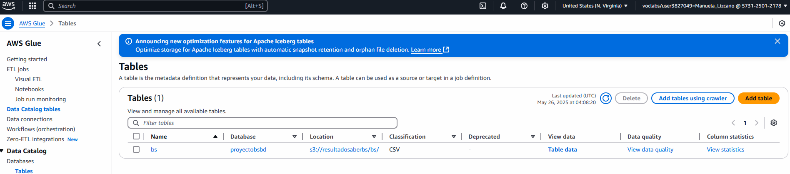


Ilustración 5-AWS\_2

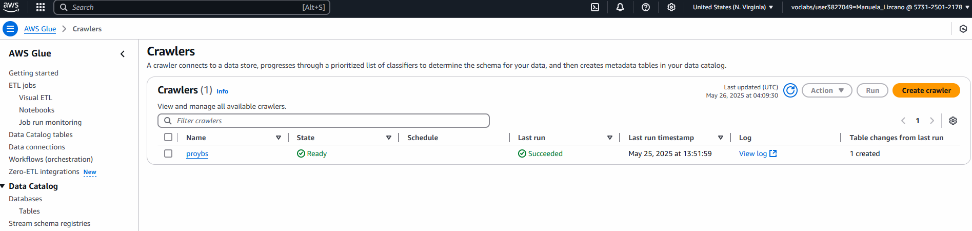


Ilustración 6-AWS\_3

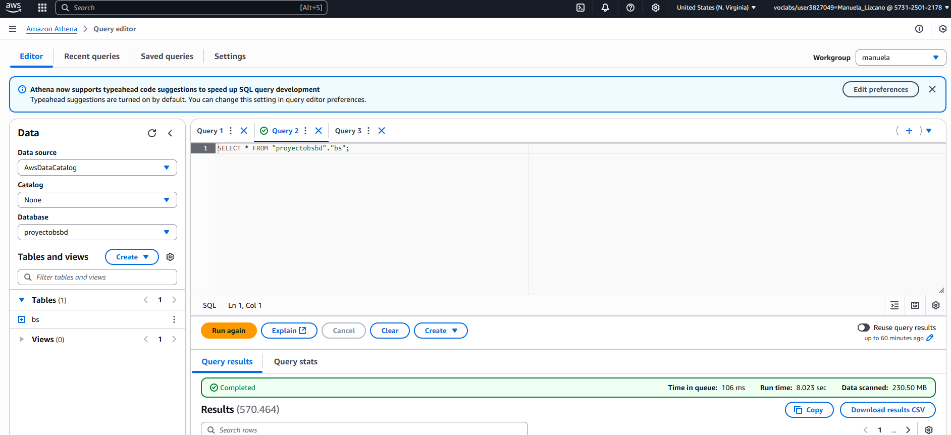


Ilustración 7-AWS\_4

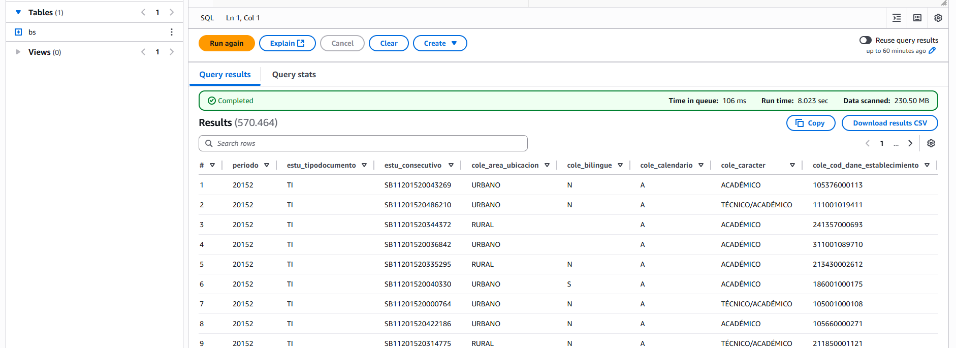


Ilustración 8-AWS\_5

Para poder realizar la limpieza de datos se comenzó con un análisis exploratorio inicial del conjunto de datos de los resultados de la prueba saber 11 del año 2015 el cual contiene **51 variables y 570,464 registros**. En primer lugar, se encontró el porcentaje de celdas faltantes lo cual dio un resultado del **0.3%** del total de datos, lo que corresponde a **86,027 celdas vacías**. Por otro lado, encontramos **2,067 datos duplicados** los cuales representan el **0.4%** del total de datos. Teniendo en cuenta esto, los datos duplicados fueron eliminados Es importante mencionar, que se encontró que el set de datos incluye **26 variables categóricas, 10 variables de tipo texto** y **15 variables numéricas.** Esta información será de utilidad para los siguientes pasos.

# Selección de Variables

Durante el proceso de preparación de datos para el modelo predictivo, se realizó una reducción significativa en la cantidad de variables disponibles. Inicialmente se contaba con un conjunto de 51 variables, las cuales fueron analizadas con Pandas Profiling. A partir de dicho análisis, se identificaron variables con alta correlación entre sí, redundancias, campos irrelevantes para la predicción que no aportaban para el modelo. Como resultado, se eliminaron 34 columnas, quedando un total de 17 variables finales para el entrenamiento del modelo.

Entre las variables eliminadas se encuentran identificadores únicos (como códigos de municipios, departamentos o instituciones) y variables de contexto no relacionadas directamente con el rendimiento académico, como el tipo de documento o el período de aplicación del examen. También se removieron los puntajes por área individual (por ejemplo, matemáticas o lectura crítica), conservando únicamente el puntaje global para evitar información reduntante. Las variables seleccionadas para permanecer fueron aquellas con relación directa a la pregunta de negocio, enfocándose en las características del entorno familiar y del colegio que podrían estar asociadas a un alto desempeño académico, lo cual permitió diseñar un modelo más claro, interpretable y alineado con los objetivos del proyecto.

Con esto en mente, se seleccionaron únicamente las variables que mostraron una relación significativa con el objetivo de negocio, alineadas con la predicción del desempeño académico a partir del contexto familiar y escolar del estudiante. Las variables finales incluidas en el modelo fueron: cole\_area\_ubicacion, cole\_bilingue, cole\_caracter, cole\_genero, cole\_mcpio\_ubicacion, cole\_naturaleza, cole\_sede\_principal, estu\_genero, estu\_mcpio\_reside, fami\_cuartoshogar, fami\_educacionmadre, fami\_educacionpadre, fami\_estratovivienda, fami\_personashogar, fami\_tieneautomovil, fami\_tienecomputador, punt\_global. Estas variables fueron consideradas como las más relevantes y predictivas para determinar la probabilidad de que un estudiante obtenga un puntaje que le permita aspirar a ingresar a la Universidad de los Andes.

# Modelos

Para construir un modelo robusto de predicción de alto desempeño en la prueba Saber 11, se utilizó un conjunto de datos previamente codificado. La variable objetivo es alto\_desempeño, una variable binaria que indica si el estudiante alcanza o no un puntaje mayor a 310. Se excluyó la columna PUNT\_GLOBAL para evitar sobre estimar el modelo. Posteriormente, los datos fueron divididos en un conjunto de entrenamiento (80%) y uno de prueba (20%) para poder realizar los modelos de entrenamiento y ver con cual se obtenía mejores resultados.

Para optimizar la estructura, se utilizó Keras Tuner con una búsqueda aleatoria de hiperparámetros, evaluando combinaciones de número de unidades por capa (16, 32, 64), funciones de activación (relu, tanh) y tamaño de capas ocultas adicionales (de 8 a 64). Cada modelo se entrenó por hasta 30 épocas, incorporando la técnica de EarlyStopping para prevenir el sobreajuste. Una vez seleccionado el mejor conjunto de hiperparámetros, el modelo fue reentrenado desde cero y evaluado con gráficas de precisión y pérdida por época para analizar su desempeño a lo largo del entrenamiento. Con esto último se logra encontrar cual es el mejor número de batches y de épocas para lo que se quiere predecir. Todo este proceso fue registrado utilizando MLflow, lo cual permitió llevar un control detallado de los parámetros, métricas y versiones del modelo propuesto.

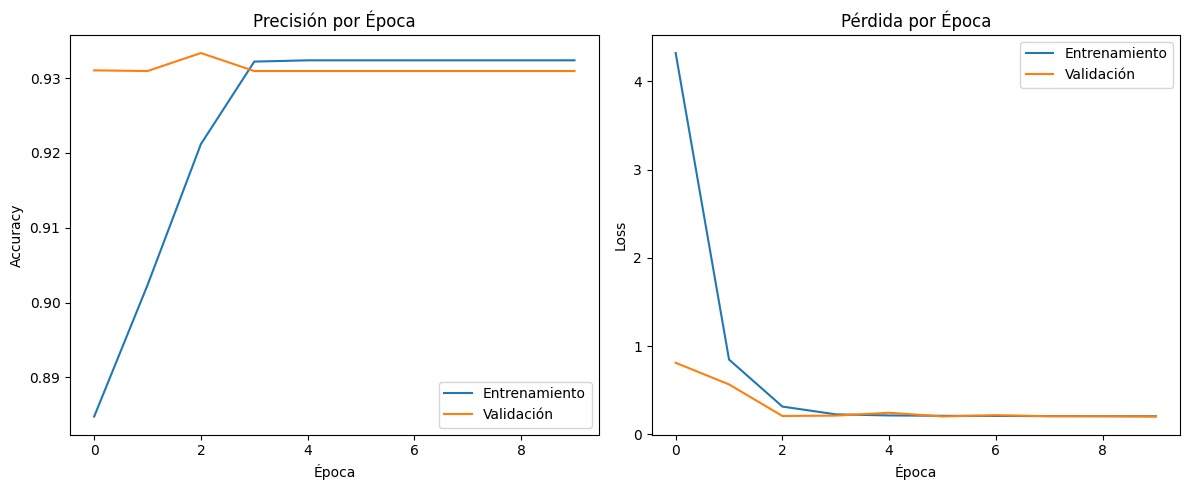
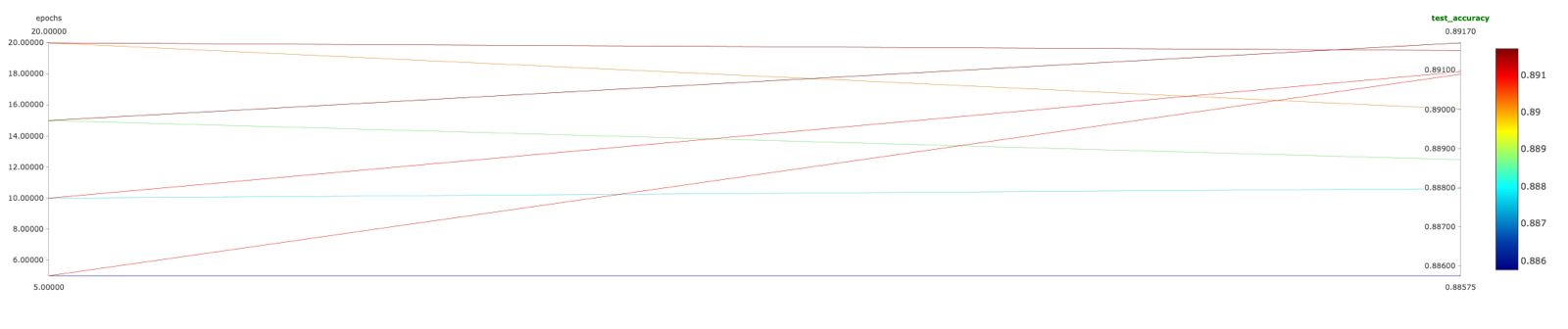
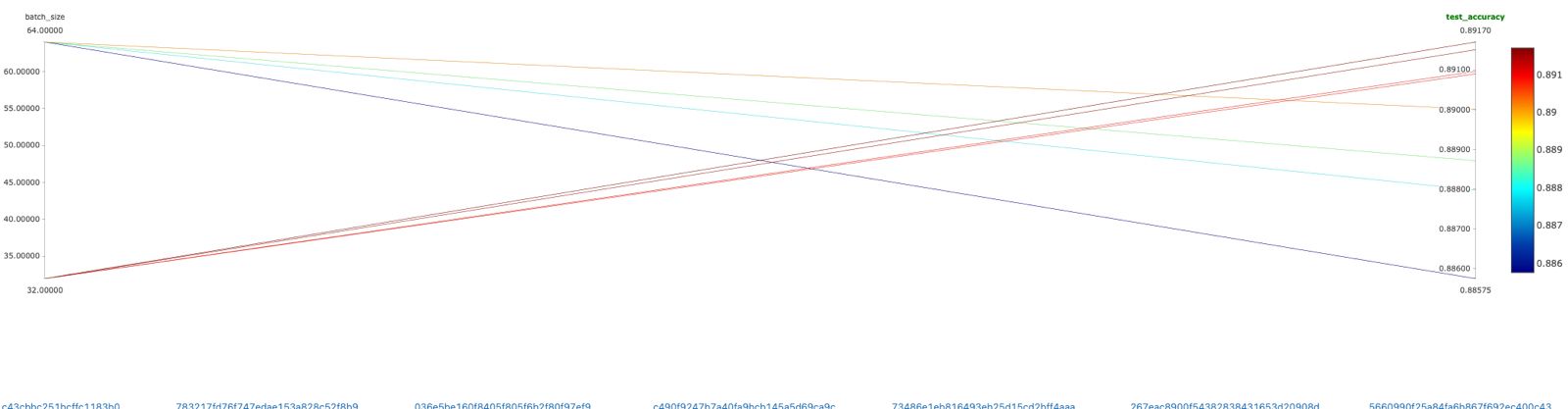
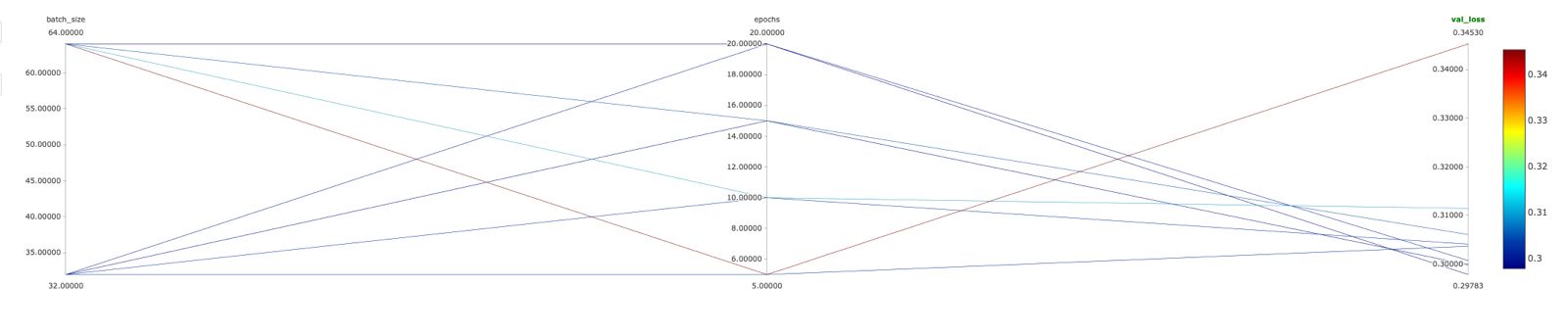


Ilustración . Gráficas de perdida y precisión



# Tableros

Se realizo un tablero interactivo el cual en primer lugar carga el conjunto de datos del examen saber 11 del año 2015 para poder realizar distintas graficas descriptivas para poder mostrar visualmente los datos de forma clara y así encontrar el impacto de las distintas variables en el rendimiento académico del estudiante. Para la primera gráfica se muestra un menú desplegable para que el usuario pueda seleccionar entre la educación de la madre o el padre, una vez seleccionado se mostrará a través de un box plot el cual ilustrara el impacto de esta variable respecto al puntaje global obtenido por el estudiante. Así mismo, se realiza un segundo gráfico el cual tiene dos botones desplegables que hacen referencia al eje X y al eje Y, en donde se encuentran disponibles todas las variables para que el usuario pueda observar mediante un gráfico de dispersión el comportamiento de las distintas variables que este escoja entre sí. Por otro lado, se carga el mejor modelo entrenado de clasificación basado en redes neuronales. Posterior a esto, nuevamente se muestran los distintos botones desplegables para poder realizar la predicción que determina si obtiene un puntaje por encima de los 310 puntos. En caso de obtener un puntaje superior a 310, se muestra un gráfico de las posibles carreras a las que puede ingresar con base al puntaje obtenido.

# Evaluación

En el proceso de evaluación se abordaron dos preguntas centrales de negocio. La primera, de tipo descriptivo, buscaba analizar si el nivel educativo de los padres influía en el desempeño académico de los estudiantes en las pruebas Saber 11. A partir del análisis exploratorio se evidenció que, en promedio, los estudiantes cuyos padres tienen un nivel educativo más alto (como educación profesional completa o posgrado) tienden a obtener puntajes más altos en el componente global de la prueba. Esto se observó tanto para la educación de la madre como del padre, reforzando la hipótesis de que el entorno académico familiar juega un rol importante en el desempeño escolar. Las visualizaciones desarrolladas en el tablero permiten al usuario explorar esta relación de forma interactiva, seleccionando distintas variables y áreas evaluadas.

La segunda pregunta, de carácter predictivo, se centró en identificar qué variables del entorno familiar y escolar estaban más asociadas con un puntaje que permita ingresar a la Universidad de los Andes. Para ello, se entrenó un modelo de clasificación basado en redes neuronales. El modelo permitió evidenciar que algunas de las variables más influyentes en la predicción de alto desempeño fueron el estrato socioeconómico, el municipio de residencia, así como la tenencia de computador y automóvil en el hogar. Estas variables reflejan condiciones sociodemográficas que parecen estar estrechamente relacionadas con mejores resultados en las pruebas. Sin embargo, esto también evidencia un sesgo del modelo o del país: las condiciones socioeconómicas, más que factores netamente académicos, están ejerciendo un fuerte impacto sobre las oportunidades de acceso a la educación superior, lo cual resalta la necesidad de políticas educativas más equitativas.

# Despliegue y mantenimiento

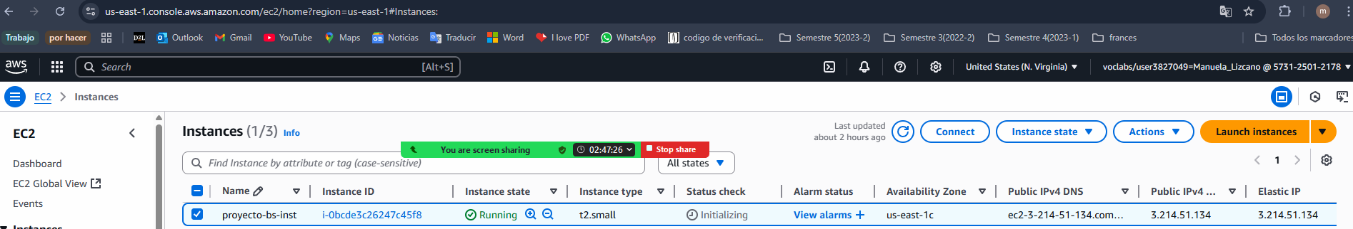
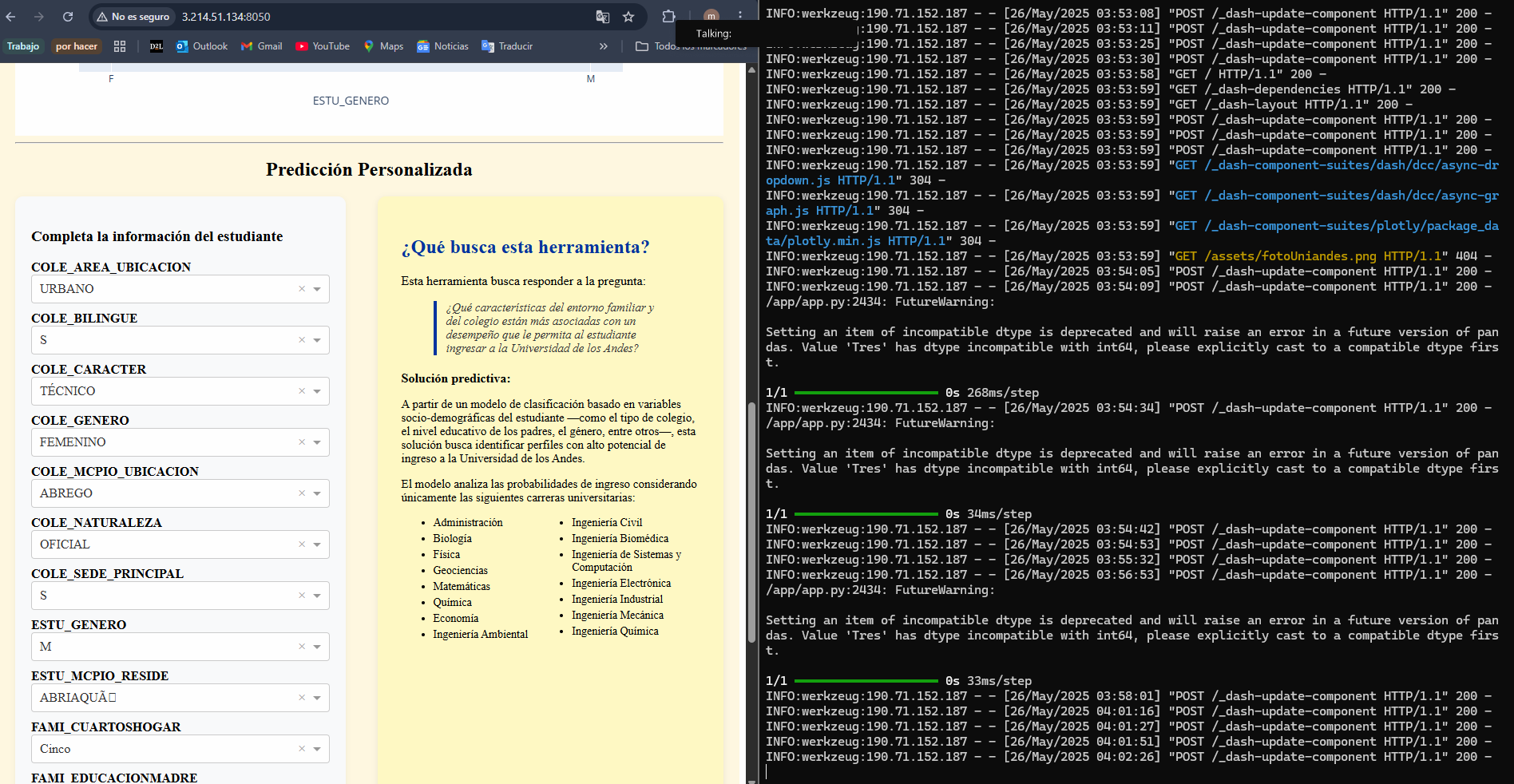


Ilustración 10-Despliegue