

Plataforma Integrada de Análisis y Consulta Inteligente para la Gestión de Datos Académicos en IMMUNE

MASTER IA & DATA SCIENCE FOR BUSINESS

OCT 24 - DIC 25



Sergio Martín Álvarez

Pablo Zafra-Polo Borrego

Tutor: Carlos E. Borges Chavez

PROBLEMA

Propuesta de mejora para el sistema de IMMUNE para analizar su actividad, el flujo de usuarios en la web, el embudo de conversión, los perfiles de los leads y los feedbacks de los alumnos para facilitar una visión global y tomar decisiones rápidas.

SOLUCIÓN

Diseñar un ecosistema completo de datos que centraliza todas las fuentes, genera métricas claras y permite analizar cursos, satisfacción y comportamiento de usuarios mediante dashboards y un agente conversacional inteligente.

FLUJO DE DATOS

Acceso desde enlaces publicitarios diversos en redes sociales

Página immune y filtro de programa

Solicitud de información y registramos el Id_usuario

Accede a los datos de programa y procede o no a matriculación

Durante la formación obtenemos los feedbacks



IMMUNE
TECHNOLOGY INSTITUTE

Click en programa

+INFO

Iniciar sesión Crear cuenta

Formulario

¿Qué te interesa

Selecciona

Nombre Apellido

Email

Número de teléfono

Spain (España) +34

Nombre de la empresa

☐ He leído y acepto el [Aviso Legal](#) y la [Política de Privacidad](#)

Enviar

Detalle programa

Gestión
matricula

Finaliza
sesión



Feedbacks de
alumnos

IMMUNE
TECHNOLOGY INSTITUTE

1

Web scraping (Open AI)

Cursos immune

2

Generación datos sintéticos

- Métricas web
- Formularios
- Feedbacks

3

Visualización datos (PBI)

Dashboard interactivo
sobre los datos de tráfico de la
web de immune

4

Análisis datos (EDA)

Transformación de datos

5

Modelo de clustering

Para los datos de tráfico de la
web de Immune y rellenan el
formulario para solicitar info

6

Agente conversacional

Sobre los datos de
feedbacks de los alumnos

OBJETIVOS DEL PROYECTO



Generación de Datos

Crear datasets sintéticos coherentes que permitan construir y validar un pipeline completo sin necesidad de datos reales



Dashboards

Diseñar un panel interactivo que centralice métricas clave y ofrezca una visión clara del rendimiento y la conversión.



Agente Conversacional

Desarrollar un agente ADK capaz de consultar cursos y analizar feedbacks en tiempo real para apoyar decisiones académicas.



Clustering

Aplicar técnicas de agrupación para identificar perfiles de leads y comprender patrones de comportamiento relevantes.



Generación de datos de forma sintética simulando un flujo completo del usuario/alumno

Métricas immune

usuario_temp: ID temporal de sesión.

origen_plataforma: Canal de adquisición (LinkedIn, Google, etc.).

IP_usuario: Dirección IP del usuario.

tiempo_en_pagina: Segundos de permanencia.

fecha_hora: Fecha y hora de la visita.

Localizacion: Ciudad y país.

programa_oferta_click: Código del curso visitado.

Id_usuario: ID único de alumno (si está registrado).

Dispositivo: Tipo de dispositivo (mobile, desktop, tablet).

Matriculado: Indicador de conversión (True/False).



Formularios

id_usuario: Identificador único del alumno (U####).

Ciudad: Ciudad de residencia.

País: País de residencia.

Edad: Edad del alumno (numérico).

Género: Género declarado (Hombre, Mujer, No contesta, etc.).

Área de interés para formarse: Sector educativo de interés.

Titulación académica: Nivel máximo de estudios completados.

Área de estudios: Campo de la titulación previa.

Experiencia laboral: Años de experiencia (numérico).

Sector laboral: Sector donde trabaja actualmente.

Motivo de la formación: Razón principal para realizar el curso.



Feedbacks

Id_encuesta: ID único de respuesta.

id_usuario: ID del alumno.

Id_curso: Código del curso evaluado.

Tipo_clase: Modalidad (Online, Presencial, Híbrido).

fecha: Fecha de la encuesta.

(Todas las variables de los cuestionarios de google)

...

velocidad_respuesta: Rapidez soporte.

utilidad_anuncios: Relevancia avisos.

satisfaccion_general: Valoración global (1-5).

comentarios: Texto libre



Cursos immune

id_curso: Código identificador del curso (C####).

nombre: Nombre oficial del programa.

tipo_de_programa: Categoría (Máster, Bootcamp, Curso, etc.).

descripcion: Breve resumen del contenido.

horas_totales: Carga lectiva total en horas.

duracion: Duración en meses/semanas.

modalidad: Formato de impartición (Online, Híbrido, Presencial).

inicio: Fecha de próxima convocatoria.

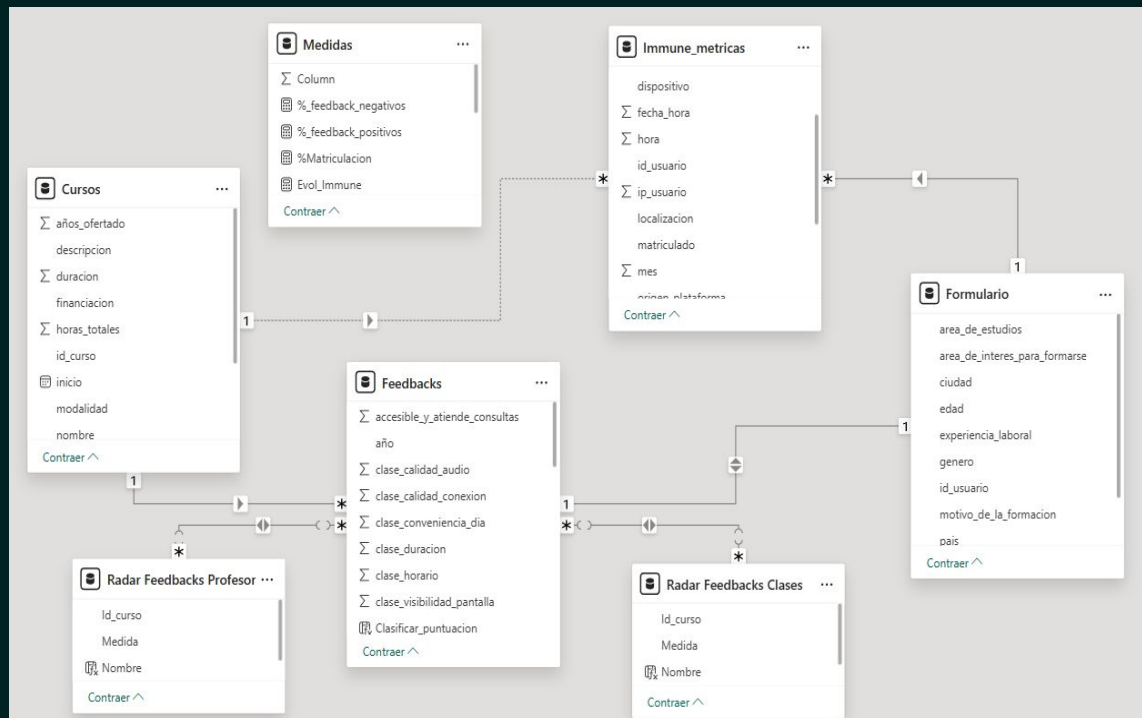
salidas_profesionales: Roles laborales a los que habilita.

financiacion: Opciones de pago disponibles.

precio: Coste del programa.

años_ofertado: Antigüedad del curso en el catálogo.





Encuestas



- ¿Cómo son las valoraciones de los programas?
- ¿Qué puntuación media obtienen?
- ¿Se mantiene durante el curso?
- ¿Son así por el profesor o por las clases?

Visitas



- ¿Cómo evolucionan durante el año?
- ¿Tenemos buen ratio de matriculación?
- ¿Tiene influencia cómo llegan a nuestra página?
- ¿Por qué se matriculan los alumnos?
- ¿Qué programas reciben más visitas?



Flujo de interacción

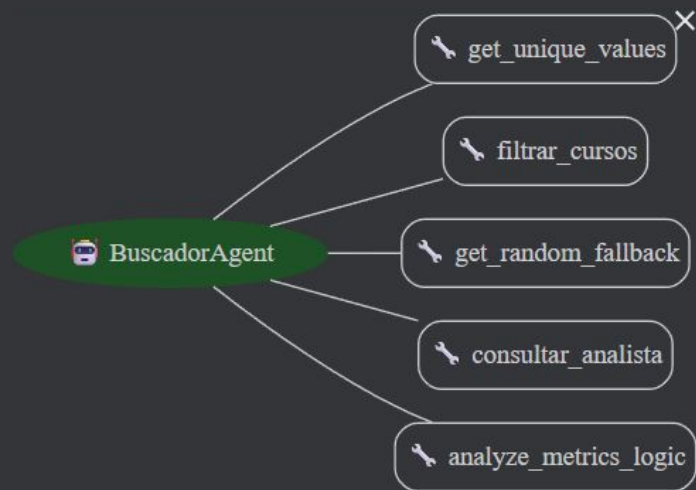
- Inicio con múltiples vías de búsqueda (sector, tipo, precio, nombre o métricas globales).
- Filtrado progresivo del catálogo mediante herramientas de consulta.
- Identificación de un curso válido.
- Generación automática de un **informe de satisfacción basado en feedbacks reales**.

Capacidades clave

- Búsqueda inteligente de cursos con criterios parciales.
- Análisis cuantitativo y cualitativo de satisfacción por curso.
- Métricas globales y rankings de la escuela.
- Gestión robusta de ambigüedades mediante sistema de fallbacks.



Agente estable desarrollado con ADK que no alucina datos, guía al usuario y genera insights accionables en tiempo real.





Herramientas del Agente

- **Exploración del catálogo** - *get_unique_values*
Muestra las opciones reales disponibles (sectores, tipos de programa, modalidades), evitando respuestas inventadas.
- **Búsqueda de cursos** - *filtrar_cursos*
Localiza cursos a partir de criterios parciales como texto libre, precio máximo o modalidad.
- **Métricas globales** - *metrics_tool*
Calcula medias generales, rankings de mejores y peores cursos y valoración del profesorado.
- **Control conversacional** - *fallback_tool*
Gestiona entradas ambiguas y mantiene el flujo de conversación estable sin alucinaciones.



Análisis de satisfacción (core del valor) - *consultar_analista*
Genera informes automáticos por curso combinando métricas cuantitativas y análisis cualitativo de comentarios.



¿Existen patrones entre las **visitas a la página** de Immune?

tabla_resumen_ip

- Recuento de registros
- Tiempo en la página
- Frecuencia de las visitas
- Programas distintos consultados
- Sectores distintos consultados
- Visitas sin acceder a programas
- Comparativas respecto a métricas por programa
- Cómo han llegado a Immune

¿Existen patrones entre los **usuarios que solicitan información**?

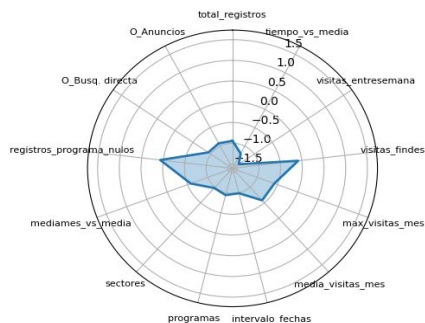
data_usuarios

- Recuento de registros
- Frecuencia de las visitas
- Programas y sectores consultados
- Motivo de la formación
- Sector laboral
- Interés en formación
- Nivel de estudios
- Área de estudios
- Edad
- Experiencia laboral

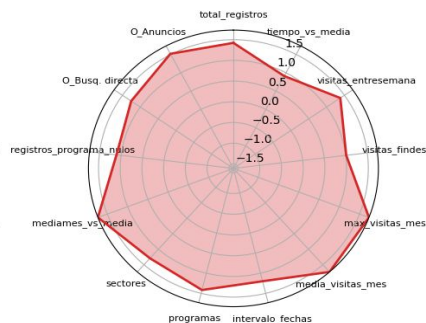




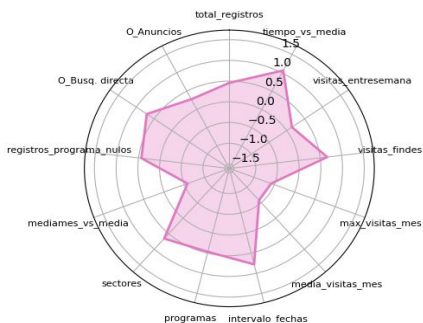
Cluster 0



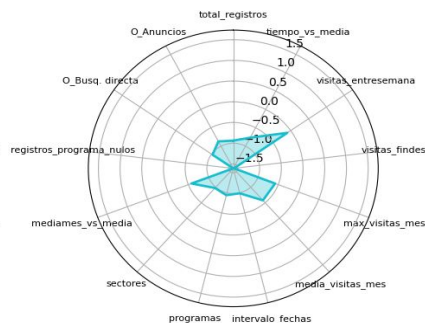
Cluster 1



Cluster 2

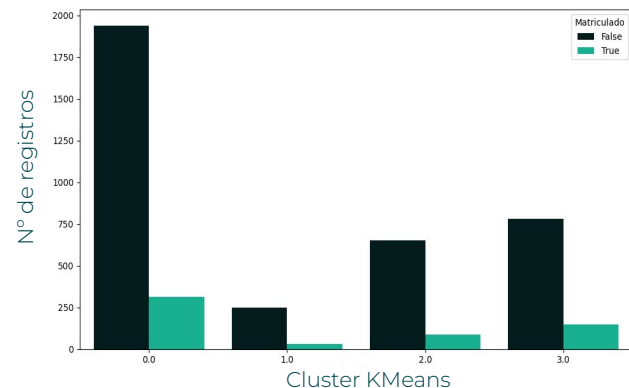


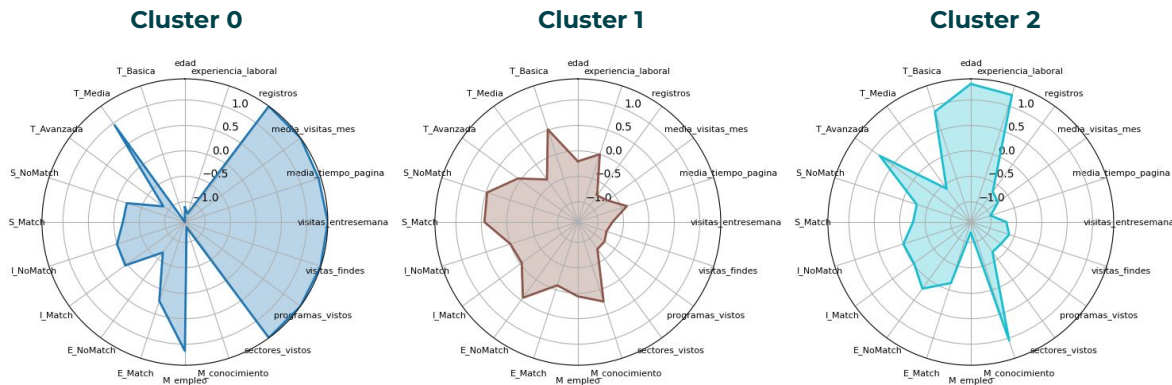
Cluster 3



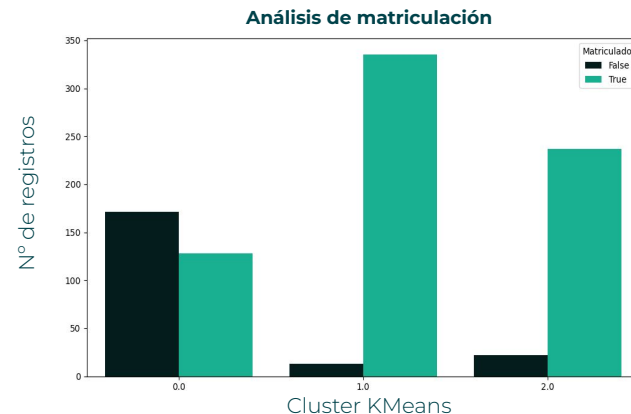
- **Cluster 0:** Escasas visitas y muy breves. No consultan programas y suelen navegar fines de semana.
- **Cluster 1:** Visitan la página muchas veces durante distintos meses accediendo por distintas fuentes y consultando gran cantidad de programas y sectores
- **Cluster 2:** Acceden con frecuencia moderada y suelen pasar más tiempo en la página aunque a veces no consultan programas
- **Cluster 3:** Bajas visitas concentradas los fines de semana influenciados por anuncios. Solo revisan un programa.

Análisis de matriculación





- **Cluster 0:** Jóvenes sin experiencia laboral que han terminado su grado o FP y buscan un primer empleo. Pasan mucho tiempo en la página consultando varios programas y sectores.
- **Cluster 1:** Personas con titulación básica, con experiencia y estudios diferentes a los ofertados en Immune que consultan pocos programas en busca de conocimiento o ascenso en su empresa.
- **Cluster 2:** Adultos con titulaciones de Máster o Doctorados y amplia experiencia laboral que consultan los programas para adquirir conocimiento.



POSIBLES MEJORAS

Establecer una estructura de Medallion en DataBricks que permita actualizar y limpiar periódicamente los ficheros de formularios de registro, visitas a la página y reseñas de los cursos en Immune.

Incluir al proyecto una tabla maestra de módulos y profesores en cada programa para profundizar aún más en valoraciones y detectar picos positivos y negativos durante todo el curso lectivo.

Aplicar el modelo de clustering de alumnos para clasificarlos en las "Casas" que después compiten en los eventos de Discord

Aprovechar las reseñas positivas y métricas obtenidas sobre las clases y profesores de las encuestas a los programas ofertados para promover la matriculación.

Elaborar nuevas consultas en el ADK y desarrollar un RAG que permitan obtener análisis globales de las reseñas filtrando por sector, tipo de programa o modalidad.

FUTUROS PROYECTOS

Utilizando los cluster detectados e información sobre campañas de marketing podríamos implementar un modelo predictivo que nos indique el porcentaje de matriculación de las visitas y el éxito de las campañas.

Con información sobre la asistencia a las clases y las encuestas de los módulos podríamos detectar los motivos por los que a final del curso baja la asistencia a las clases.

Con información sobre el proceso de matriculación y sus fases, podríamos implementar otro modelo de cluster para encontrar patrones entre los alumnos que finalizan el proceso matriculándose en algún programa y los que no.

Implementar un agente en la página de Immune para resolver dudas sobre los programas e incluso pueda dar información sobre los profesores a partir de sus perfiles de LinkedIn.



Estimación de Costes – Agente con API Gemini

Batch de Análisis (semanal / mensual / trimestral)

Análisis automático de datasets grandes (~20.000 filas)

- Modelo: Gemini 2.5 Flash
- Coste por batch \approx 0,30 – 0,35 USD

Coste estimado

- Semanal (4/mes): \sim 1,3 USD / mes
- Mensual: \sim 0,35 USD / mes
- Trimestral: \sim 0,35 USD / trimestre

RAG con Embeddings Profesionales

Indexado y consulta semántica de feedbacks a gran escala

- Modelo de embeddings: Gemini Embedding
- Coste embeddings \approx 0,10 USD / millón de tokens

Coste Estimación

- Indexado inicial (\approx 1–2 M tokens): \sim 0,10 – 0,20 USD (one-off)
- Consultas RAG (por interacción): $<$ 0,01 USD



Costes muy bajos y totalmente viables para producción: el sistema permite análisis periódicos y RAG avanzado con coste mensual inferior a 5 USD en escenarios realistas.