TFM DEFENSA COMILLAS ICADE 2025-2026

Sistema de Lead Scoring y Asistente Virtual con RAG para la captación de estudiantes en la Universidad Pontificia Comillas

AUTOR

Pedro Pérez Blanco

TUTOR

José Portela González

MÁSTER

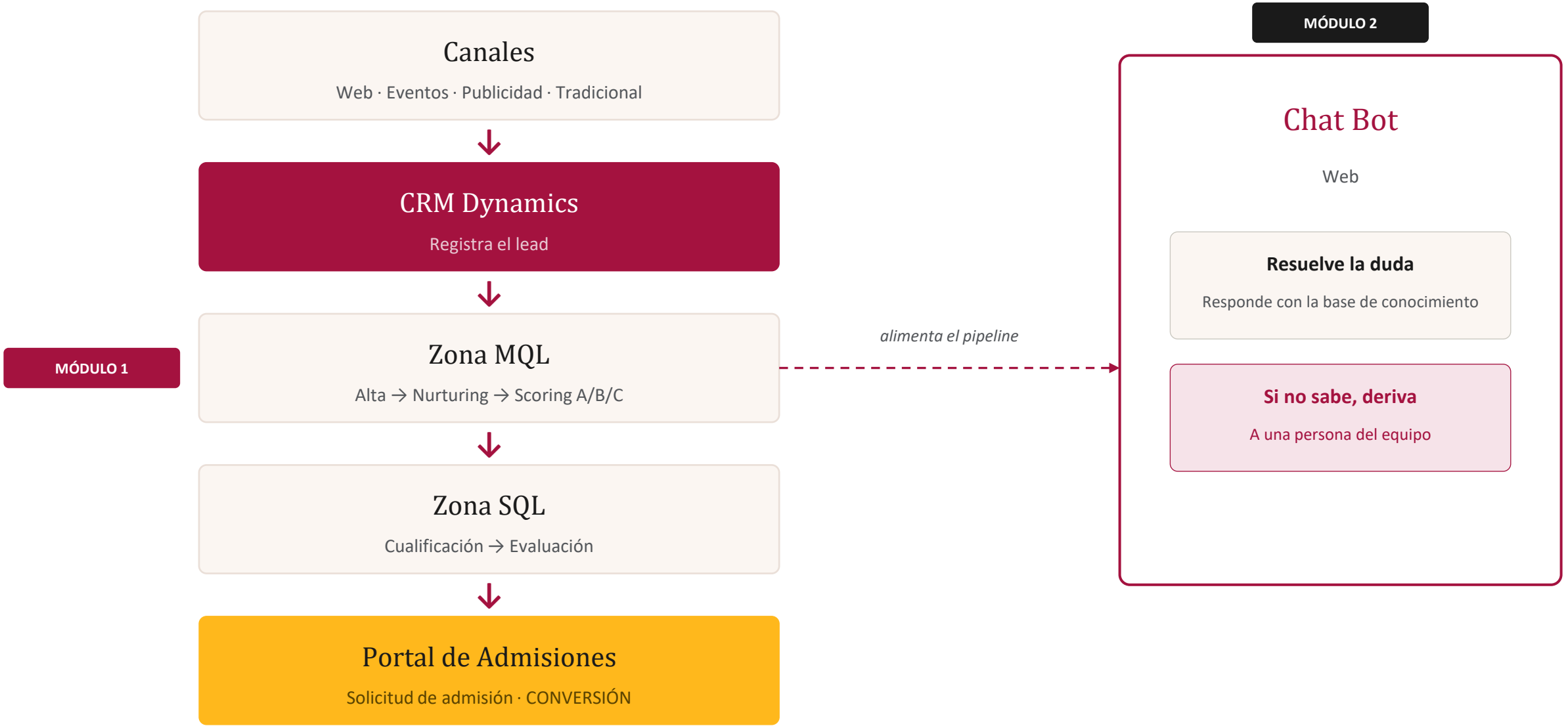
Business Analytics

CONVOCATORIA

Junio 2026

Índice

01 El recorrido del lead, de principio a fin



03 Un sistema integrado en dos módulos

Priorizar a quién contactar y automatizar la respuesta a las dudas. Dos módulos independientes que se apoyan en un protocolo común de derivación.

MÓDULO 1

Lead scoring

Modelo de clasificación supervisada (Random Forest) que estima la probabilidad de que cada lead entregue documentación y la traduce en categorías A, B o C para priorizar el contacto.

Random Forest · scikit-learn

MÓDULO 2

Asistente RAG

Chatbot con arquitectura RAG sobre Microsoft Copilot Studio: recupera la información de una base de conocimiento de Comillas y genera la respuesta solo a partir de ella.

Copilot Studio · base de conocimiento

TRANSVERSAL

Fallback

Protocolo de seguridad: cuando el bot no puede resolver la consulta, recoge el contacto y la deriva a un asesor humano, convirtiendo la duda en un lead.

Derivación a una persona

Tres metas medibles

1 **A + B capturan ≥ 95%**
de las entregas de documentación

2 **El bot resuelve solo**
las dudas frecuentes, sin intervención humana

3 **Ninguna duda se pierde**
una pregunta sin respuesta se vuelve un lead

04 Lo que ya existía y el hueco que cubro

Las tres piezas existen por separado; mi trabajo las integra. Revisé cada área para situar la aportación.

Lead scoring

Qué hay

Consolidado en banca y e-commerce para priorizar clientes.

Qué falta

Apenas aplicado a la captación universitaria.

Chatbots universitarios

Qué hay

Existen asistentes de admisión y atención al alumno.

Qué falta

La mayoría son basados en reglas; pocos recuperan la documentación real de la institución (RAG).

Protocolos de derivación

Qué hay

Algunos sistemas escalan a un agente humano.

Qué falta

Funcionan aislados, sin conectarse a la puntuación del lead.

Mi aportación. Integro las tres piezas —scoring, RAG y fallback— en un único sistema para la captación universitaria, sobre la infraestructura que Comillas ya tiene. Algo que en la literatura no aparece junto.

05

Del CRM al dataset de modelado

Cuatro decisiones metodológicas antes de modelar. Marcan la fiabilidad de todo lo que viene después.

1 Variable objetivo

Predigo la entrega de documentación, no la matrícula: las notas son confidenciales y la documentación depende solo del candidato.

2 Exclusión del Portal

Dejo fuera el Portal de Admisiones: convierte casi al 100% y distorsionaría la comparación.

3 Control de fuga

Reviso variable a variable y elimino todo lo posterior al desenlace para evitar data leakage.

4 Privacidad (RGPD)

Datos anonimizados. El sistema asiste al equipo de admisiones; no decide la admisión.

EL DATASET

24.365

leads tras los filtros, de 30.593 brutos (44 columnas)

CRM Microsoft Dynamics 365 · curso 2025-26

44 → 13 → 72 columnas

selección + one-hot

clase positiva minoritaria

desbalance fuerte

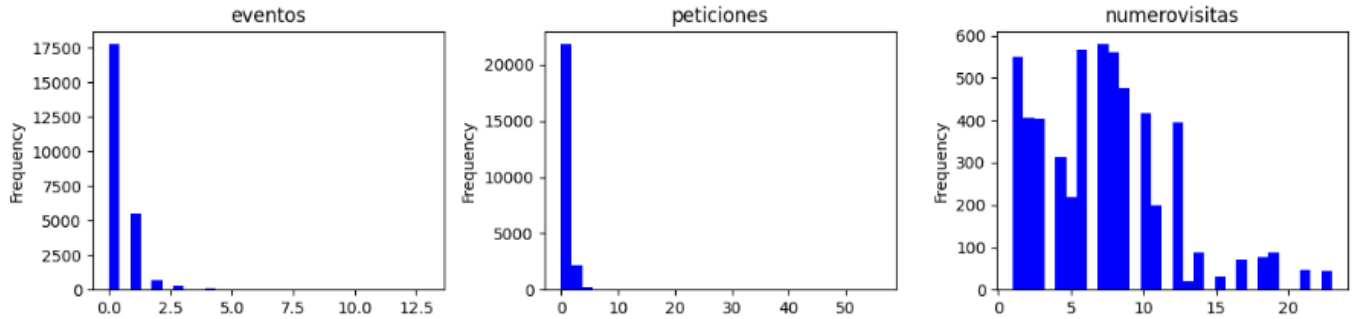
Split 80 / 20

estratificado

06

Análisis exploratorio de datos

Antes de entrenar, los datos ya dejan ver qué separa a los leads que entregan.



Engagement

Distribución de eventos, peticiones y visitas. Los leads que más interactúan con la universidad concentran las entregas: es la señal que mejor los separa.

07

Entrenamiento y evaluación de modelos

Comparo tres modelos en igualdad de condiciones, con decisiones pensadas para un dato desbalanceado (clase positiva minoritaria).

01

Preparación de variables

- De 13 predictores a una matriz de 24.365 × 72 columnas (one-hot).
- Variables creadas: zona desde el código postal, periodo de campaña desde la fecha, y agrupación de congregaciones y estudios.
- One-hot para no imponer un orden falso a las categorías.

02

Tres modelos comparados

- Regresión logística escalada, como modelo de referencia interpretable.
- Random Forest y Gradient Boosting como candidatos, con hiperparámetros ajustados por Random Search.

03

Decisiones metodológicas

- Sin SMOTE: no genero leads sintéticos → class_weight = balanced.
- Métrica principal PR-AUC, no accuracy (engaña con 10% de positivos).
- Validación 5-fold estratificada: mantiene la proporción de positivos en cada pliegue.

08

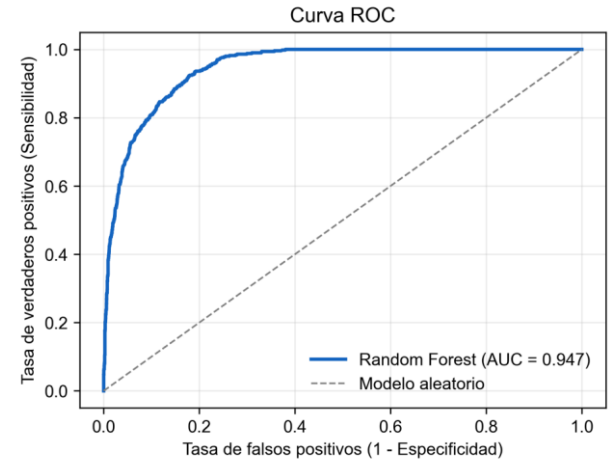
Resultados en test

Sobre datos no vistos al entrenar, gana el Random Forest. Y la caída de validación cruzada a test descarta sobreajuste.

Modelo	PR-AUC (CV)	PR-AUC (test)	Caída
Regresión logística	0,577	0,564	0,013
Random Forest	0,698	**0,691**	**0,007**
Gradient Boosting	0,695	0,685	0,010

Caída = diferencia PR-AUC entre validación cruzada y test.

Elijo Random Forest. Gana en PR-AUC (0,691) y el Gradient Boosting queda a solo 0,006. Me decido por el RF porque cae apenas 0,007 de validación a test y su importancia de variables es directamente interpretable.



Curva ROC del Random Forest sobre el conjunto de test.

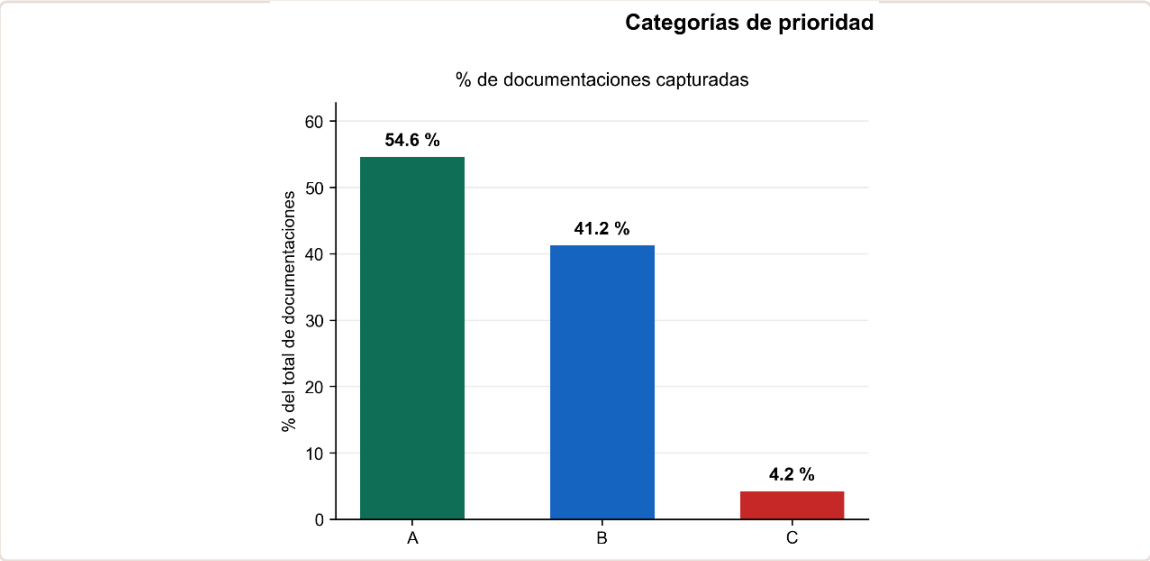
0,947	0,767	0,628
ROC-AUC	Sensibilidad	F1

Sensibilidad y F1 calculadas con el umbral fijado en entrenamiento, no en test.

09

Definición de categorías A, B y C

Una probabilidad de 0 a 1 no es operativa para admisiones. La convierto en tres categorías, con los cortes fijados sobre el conjunto de entrenamiento.



Captura de entregas por categoría — test.

Cat.	Volumen	Lift	Captura
A	8%	**6,77**	**54,6%**
B	22%	1,85	41,2%
C	70%	0,06	4,2%

Lift = veces sobre la conversión media de la base. Captura = % del total de documentaciones que cae en la categoría.

A + B: con el 30% del volumen capturo el **95,8%** de las entregas.

En la práctica. El equipo trabaja solo 1 de cada 3 leads (A y B) y aun así no pierde casi ninguna entrega. Esa era la primera meta del trabajo, y se cumple.

10 Interpretabilidad del modelo

Para que el modelo no sea una caja negra, lo interpreto por tres métodos. Si los tres coinciden, es señal de que aprende patrones reales, no ruido.

01 Importancia del Random Forest

El propio modelo mide cuánto usa cada variable para separar los casos. Da el peso global de cada variable.

02 Regresión logística

Sus coeficientes con signo dan la dirección: indican si cada variable empuja a favor o en contra de que el lead entregue.

03 SHAP (TreeExplainer)

Sobre 1.000 leads de test, calcula cuánto suma o resta cada variable a cada predicción concreta. Lectura local y global.

Qué hace SHAP

Sobre 1.000 leads del conjunto de test, TreeExplainer reparte cada predicción entre las variables que la forman: cuánto suma o resta cada una, y en qué dirección.

Por qué no está el gráfico

El diagrama de la defensa etiqueta zonas y canales de captación concretos. Eso es información interna de la universidad y no entra en la versión pública.

Qué sí se puede decir

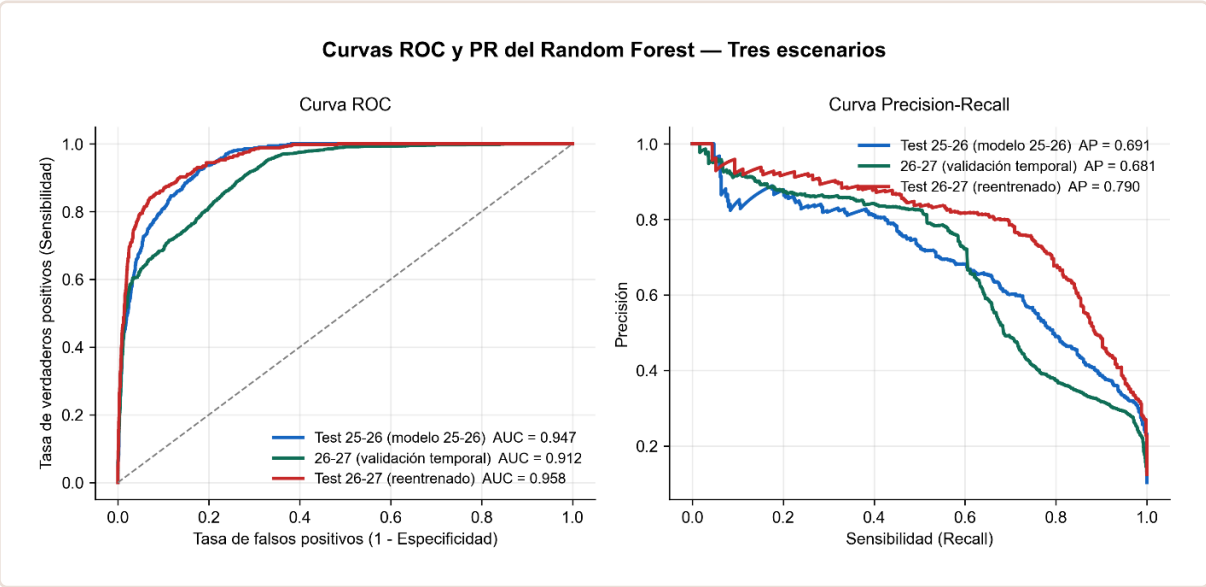
Los tres métodos, por caminos distintos, señalan las mismas familias de variables. La coincidencia es el resultado; el detalle por categoría se queda fuera.

La interpretabilidad se resume en texto: el diagrama completo no entra en la versión pública.

Coinciden los tres. Por caminos distintos, los tres métodos señalan las mismas variables —engagement, zona y canal—, las mismas del análisis exploratorio. El modelo aprende señales reales, no ruido.

11 Validación temporal

¿Se sostiene el modelo ante datos de un curso futuro? Lo aplico al curso 2026-27, posterior a los datos de entrenamiento.



Curvas ROC y Precision-Recall en los tres escenarios sobre el curso 2026-27.

Punto de partida — test 25-26: PR-AUC 0,6911

Escenario en 2026-27	PR-AUC
RF sin reentrenar	**0,6806**
RF reentrenado	0,7900
Regresión logística reentrenada	0,7604

Cada escenario se evalúa sobre su propio conjunto de test.

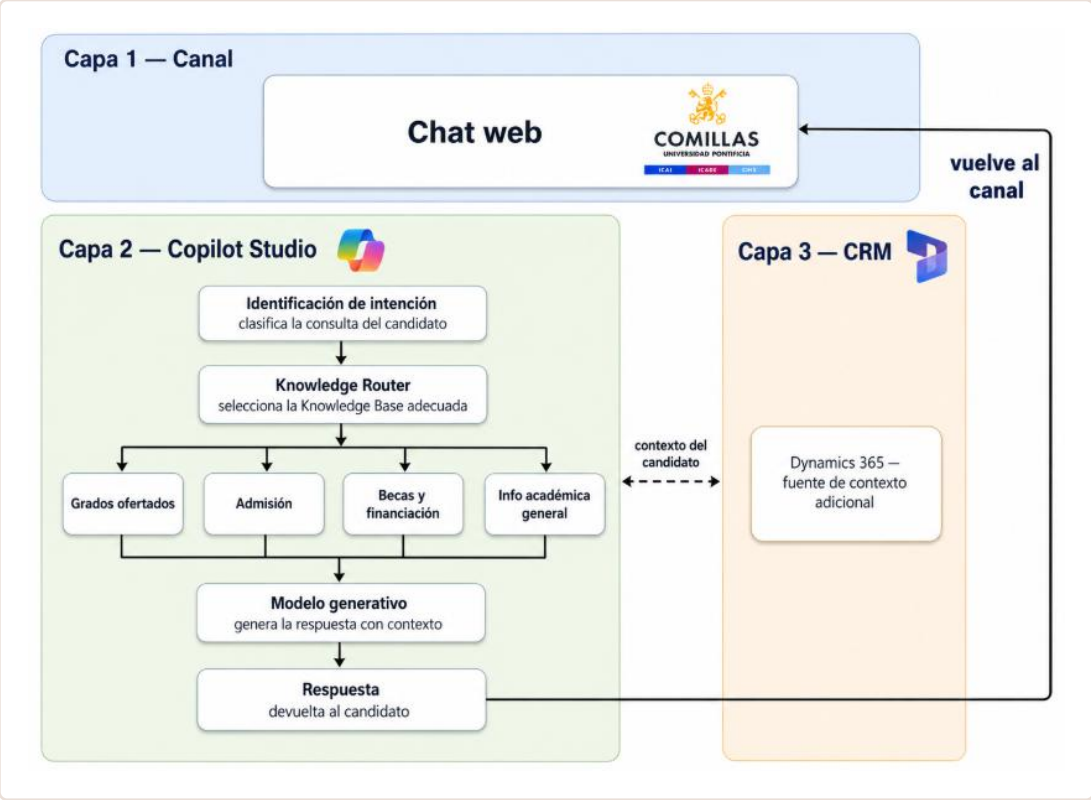
¿Se degrada con el tiempo? No: sin reentrenar cae solo 0,0105.

¿Mejora si lo actualizo? Sí: reentrenado sube a 0,7900.

Recomendación. El modelo se mantiene estable entre cursos, pero se recomienda reentrenarlo anualmente y recalibrar el umbral de la categoría B, el más sensible a variaciones en la tasa de positivos

12 Asistente virtual con RAG

El segundo módulo responde solo con la información de Comillas. Montado sobre Microsoft Copilot Studio, la herramienta corporativa del ecosistema Microsoft que ya usa Comillas.



Arquitectura por capas. La capa CRM (Dynamics 365) es prevista, no implementada en este TFM.

1. **Base de conocimiento**
   Cuatro áreas: grados, admisión, becas e información académica.

2. **Enrutador**
   Envía cada pregunta al archivo que corresponde, para no mezclar respuestas de áreas distintas.

3. **Recuperación + generación**
   Recupera la información de la base correcta y genera la respuesta solo con ella.

4. **Temas**
   Las cuatro áreas de conocimiento, más el flujo de fallback cuando no hay respuesta.

Garantía de fiabilidad. El system prompt obliga al bot a responder solo con la base de conocimiento: no inventa ni busca por su cuenta.

13 Protocolo de fallback y líneas de futuro

Cuando el bot no sabe, no improvisa: deriva a una persona. El caso de uso es deliberadamente sencillo, pero la arquitectura está pensada para crecer.

Protocolo de fallback del asistente. Elaboración propia.

IMPLEMENTADO

El bot detecta que no puede resolver la consulta, recoge los datos de contacto del candidato y ofrece que un asesor del equipo le llame. Una duda sin respuesta se convierte en un lead, no en un abandono.

Probado en el panel de Copilot Studio. Las tasas de resolución autónoma y de fallback quedan definidas, pendientes de medir en producción con candidatos reales.

LÍNEAS DE FUTURO

WhatsApp

Añadir el canal vía Azure Communication Services.

Bookings + caso automático

Agendar citas y crear el caso en Dynamics con el contexto de la conversación.

Agente proactivo

Conectado al CRM: mensajes según la fase del candidato.

Agentes con MCP

Actuar en nombre del candidato: consultar solicitud, iniciar trámites.

14 Limitaciones

Estas son las principales limitaciones del trabajo y su alcance.

1	Un solo curso de datos	Entreno con un único curso; la validación temporal ayuda, pero no equivale a disponer de varios años.
2	Sin notas académicas	Por confidencialidad no tengo el expediente; por eso predigo entrega de documentación, no matrícula.
3	Sesgos en los datos	El modelo refleja la historia de Comillas: hay sesgo geográfico y de congregación que vigilar en producción.
4	Asistente sin probar en real	El chatbot solo se ha validado en el panel de Copilot Studio, no con candidatos reales.
5	Módulos aún sin conectar	El scoring y el asistente funcionan en paralelo, pero todavía no se comunican a través del CRM.

Y por eso el sistema asiste al equipo, no decide: la última palabra siempre es humana.

15 Conclusiones

Tres resultados resumen lo que aporta el lead scoring. Todos sobre el conjunto de test, con la variable objetivo de entrega de documentación.

/ 01 COBERTURA

95,8%

de las entregas de documentación se capturan trabajando solo las categorías A y B.

/ 02 LIFT

6,77

veces más probable de entregar que la media: eso es un lead de A.

/ 03 PRIORIZACIÓN

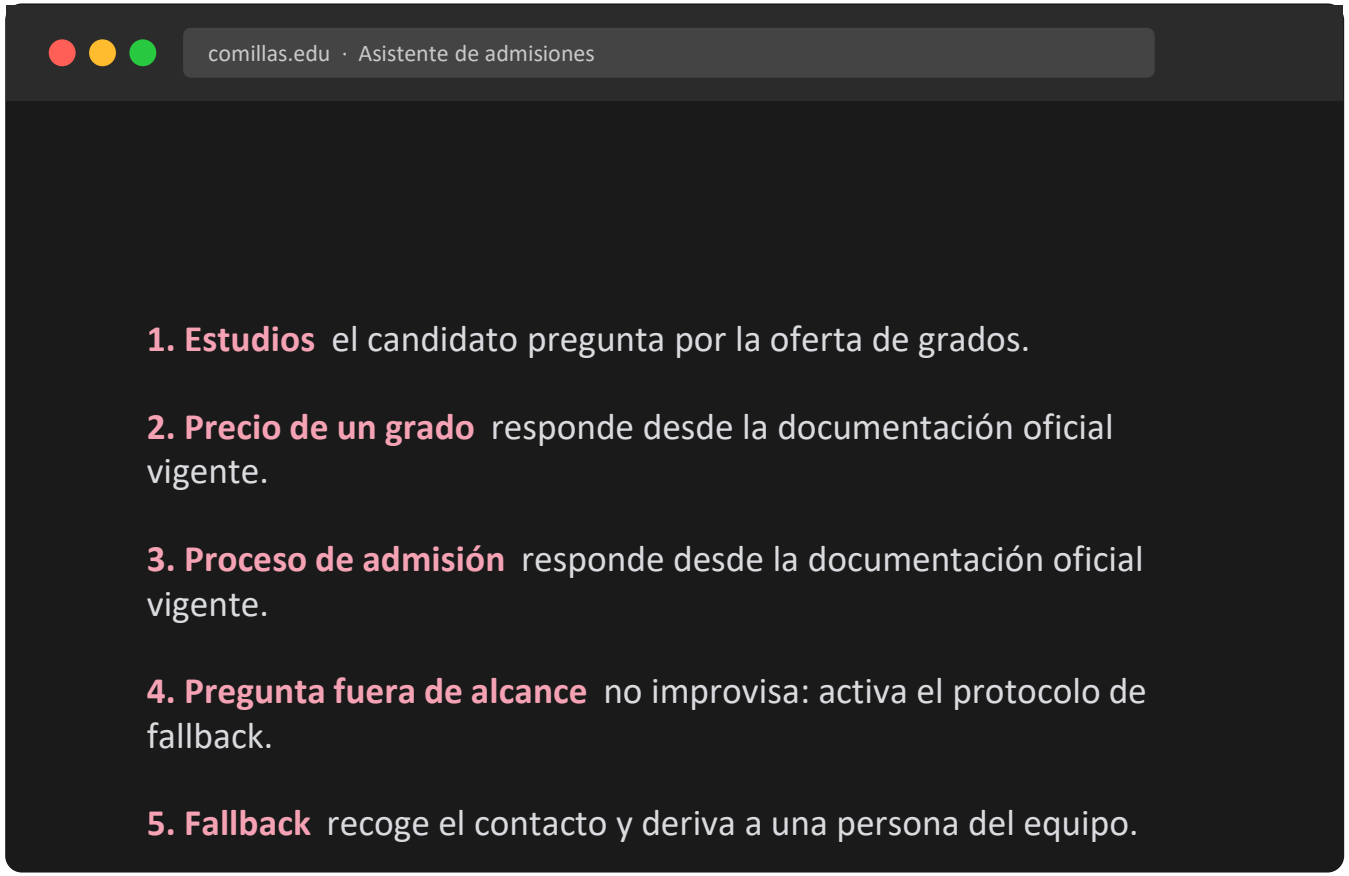
1 de 3

leads basta para no perder casi ninguna entrega.

En la práctica. Un sistema que el equipo de admisiones puede usar desde el primer día, sobre las herramientas que Comillas ya tiene, sin infraestructura nueva.

16 Demostración del asistente

La demostración se hizo en directo en la defensa. Esta versión pública no incluye la grabación, así que aquí queda el recorrido que se enseñó, paso a paso.



Estudios → precio de un grado → proceso de admisión → pregunta fuera de alcance → fallback a una persona

Gracias.

AUTOR

Pedro Pérez Blanco

CONTACTO

pedro.perez.blanco@outlook.es

REPOSITORIO

github.com/ppblanco/admissions-lead-scoring-rag