PROYECTO FINAL · APRENDIZAJE DE MÁQUINA 1INF02 · PUCP 2026-2

¿Con cuánto retraso llegará el bus al final de la ruta?

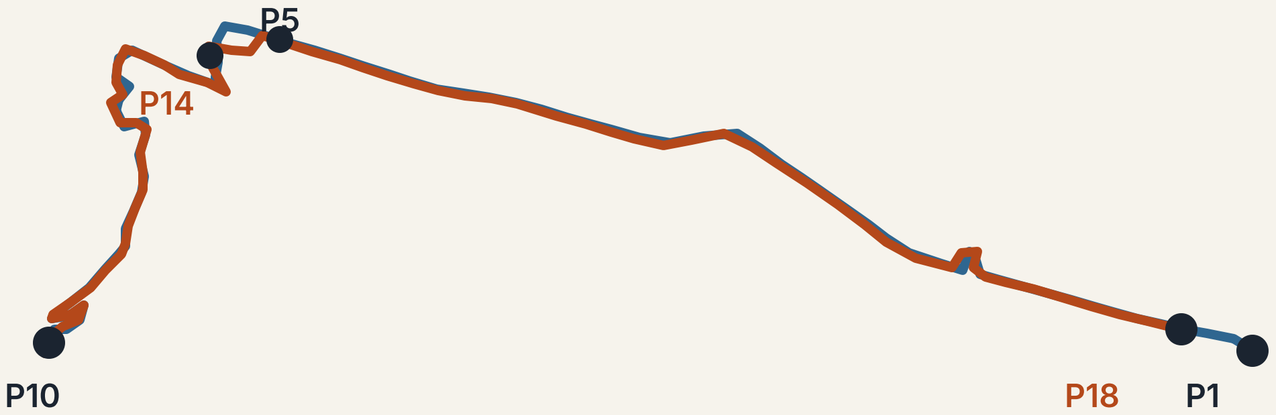
Predicción del retraso en el último punto de control de una ruta de Cusco con datos de control de horarios, GPS y aprendizaje automático

Claudia Serpa · Nicoll Tone · Reynaldo Mamani · Jeremy Aldama

27 de noviembre de 2026

INTRODUCCIÓN

El retraso aparece en la segunda mitad de la ruta



Ida P1→P10 (hacia el oeste) Vuelta P10→P18

Una sola ruta circular de 18 puntos de control con horario fijo, 45 buses, 6 ene–1 feb 2025. Ruta reconstruida con el GPS.

38 %

de las vueltas llega a P18 más de 3 min tarde; el 17 % llega 2 min o más adelantada

3,7 min

retraso mediano en P14, tras oscilar entre −1,6 y 1,6 min hasta P13

10,8 min

retraso mediano en P18 los sábados, frente a 1,1–2,5 min los otros días

INTRODUCCIÓN

Objetivo: anticipar el retraso en P18 desde P10

Cuando el bus pasa por P10, con 8 puntos por recorrer, predecir su retraso en P18 en minutos y traducirlo en una decisión.

1. Limpiar los datos con reglas explícitas y cuantificadas
2. Comparar Ridge, Random Forest y HistGradientBoosting con 3 líneas base
3. Medir cuánto aportan variables en tiempo real (otras vueltas y GPS)
4. Medir el error según el punto desde el que se predice
5. Clasificar la puntualidad y agrupar perfiles de retraso
6. Construir una demo interactiva

METAS FIJADAS ANTES DE LA PRUEBA

Superar a la mejor línea base en MAE

El MAE en minutos es la métrica principal

$R^2 \geq 0{,}80$

Umbral de referencia: explicar el 80 % de la varianza

TRABAJOS RELACIONADOS

Los ensambles de árboles son la referencia en datos tabulares

Trabajo	Datos y modelos	Lo que tomamos
Martyanov et al. 2026	AVL de Varsovia; RF, XGBoost, CatBoost	Retraso = llegada real − programada; MAE ≈ 3,6 min
Petersen et al. 2019	Línea de Copenhague; ConvLSTM	MAE 1,99 min frente a 3,23 de la media histórica
Achar et al. 2020	GPS de buses; filtro de Kalman espacial	El estado de un tramo informa sobre el siguiente
Degeler et al. 2021	AVL; k-means con silueta	Estados de operación: normal, retrasado, apelotonado
Carcausto et al. 2024	GPS de buses de Cusco; LSTM	Predice tiempo de viaje, no retraso respecto al horario
Traverso et al. 2025	Corredor Metropolitano, Lima; RF	R^2 = 0,9998 con partición 70/30: riesgo de fuga

> Brecha: no encontramos trabajos que predigan el retraso en puntos de control de Cusco ni que combinen el registro de horarios con el GPS de los buses precedentes.

METODOLOGÍA · DATOS Y LIMPIEZA

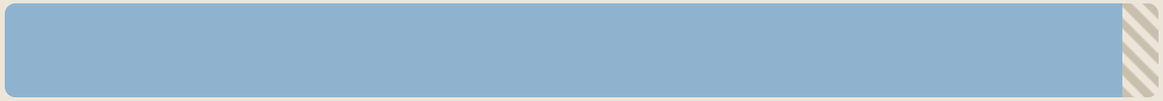
4 004 de 4 707 vueltas sirven para predecir

Vueltas que quedan tras cada regla de limpieza

Registradas en **Times.csv** · 45 buses, 18 puntos de control

4 707

Con horario programado válido · **−148** con el horario corrupto

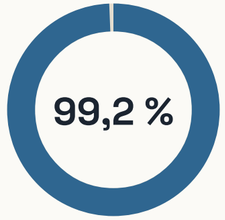


4 559

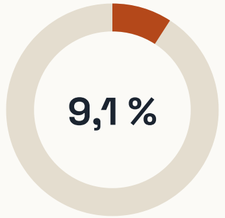
Con P18 observado · **−555** sin P18: el objetivo nunca se imputa



4 004



de los pasos se une a su GPS



de los pasos sin registrar: quedan vacíos

156

pasos con hora imposible: vacíos

55

saltos de más de 15 min: se conservan

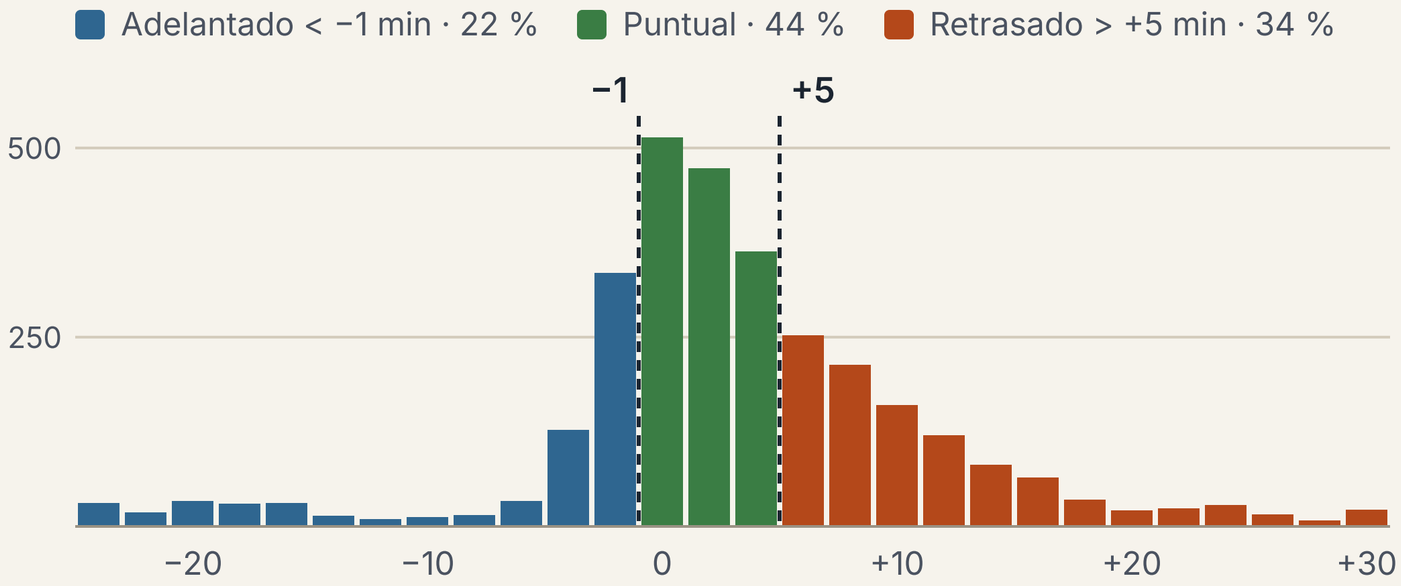
Retraso de buses en Cusco · Monitoring.csv: 1,63 M posiciones GPS · reglas fijadas antes de modelar

METODOLOGÍA · VARIABLE A PREDECIR

Qué predecimos y por qué: el retraso en P18

d18 = hora real en P18 − hora programada en P18

en minutos con segundos; positivo = tarde, negativo = adelantado



Lo que se controla y sanciona

La llegada sale de horario + retraso

P18: el retraso ya se acumuló

Mediana 2,4 min, p90 12,8 min; desde P10 quedan 8 puntos

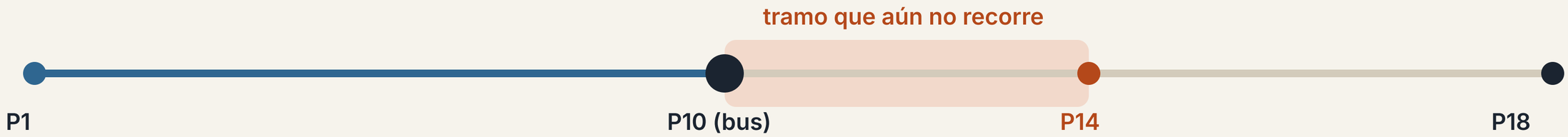
Con segundos y sin imputar

difference_time estaba truncado; solo vueltas con P18

Retraso de buses en Cusco · retraso en P18 (min) de las 3 082 vueltas del desarrollo (6–26 ene), barras de 2 min

METODOLOGÍA · VARIABLES

En P10 solo se usa lo que ya pasó



Del registro de horarios (20)

Retrasos d1…d10, último retraso y su variación, pasos faltantes, hora de salida (seno y coseno), plantilla, tiempo restante, día, vuelta y vehículo.

Tiempo real: otras vueltas

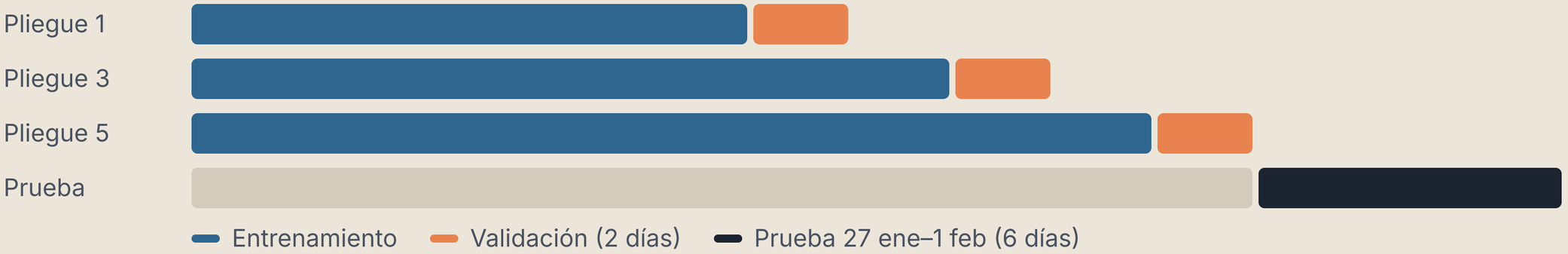
Retraso que ganaron entre P10 y P14 las vueltas que pasaron por P14 en los últimos 30 min. Entra: baja el MAE en 3 de 3 modelos.

Tiempo real: GPS adelante

Velocidad de los buses que en los últimos 15 min iban entre P10 y P14. Entra: 2 de 3. El GPS del propio bus no (1 de 3).

METODOLOGÍA · DISEÑO EXPERIMENTAL

Validación por días y prueba única



Modelos

Líneas base: retraso cero, persistencia y media por hora. Ridge, Random Forest y HistGradientBoosting con búsqueda aleatoria de hiperparámetros.

Métricas

MAE (principal), RMSE, R^2 y % con error ≤ 2 min. IC 95 % por bootstrap por día; el día es la unidad de comparación.

Prueba única

Huella SHA-256 del código registrada antes de evaluar; el script se niega a repetir la evaluación.

RESULTADO PRINCIPAL · PRUEBA 27 ENE–1 FEB, 922 VUELTAS

2,76 min

de error medio al predecir el retraso en P18 desde P10

−36 %

frente a la persistencia (4,32 min); gana los 6 días

R² 0,76

IC 95 % por día: 0,62–0,84

50 %

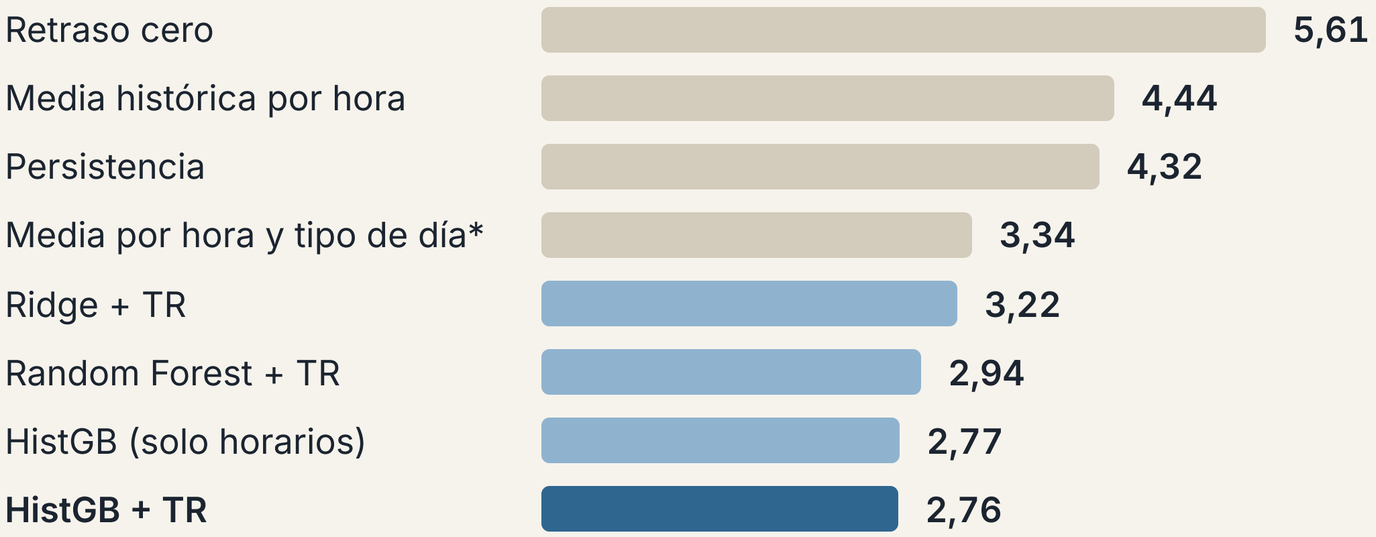
de predicciones con error de 2 min o menos

HistGradientBoosting con variables en tiempo real · IC 95 % del MAE: 2,28–3,34

RESULTADOS · COMPARACIÓN DE MODELOS

Los ensambles ganan a las líneas base

MAE en la prueba (min). TR: con variables en tiempo real



*Añadida en la revisión del E2, después de la evaluación única.

6 de 6

días gana a la persistencia: +1,56 min [0,90; 2,94]

+0,18

min sobre Random Forest + TR [−0,06; 0,37]: no concluyente

−0,46

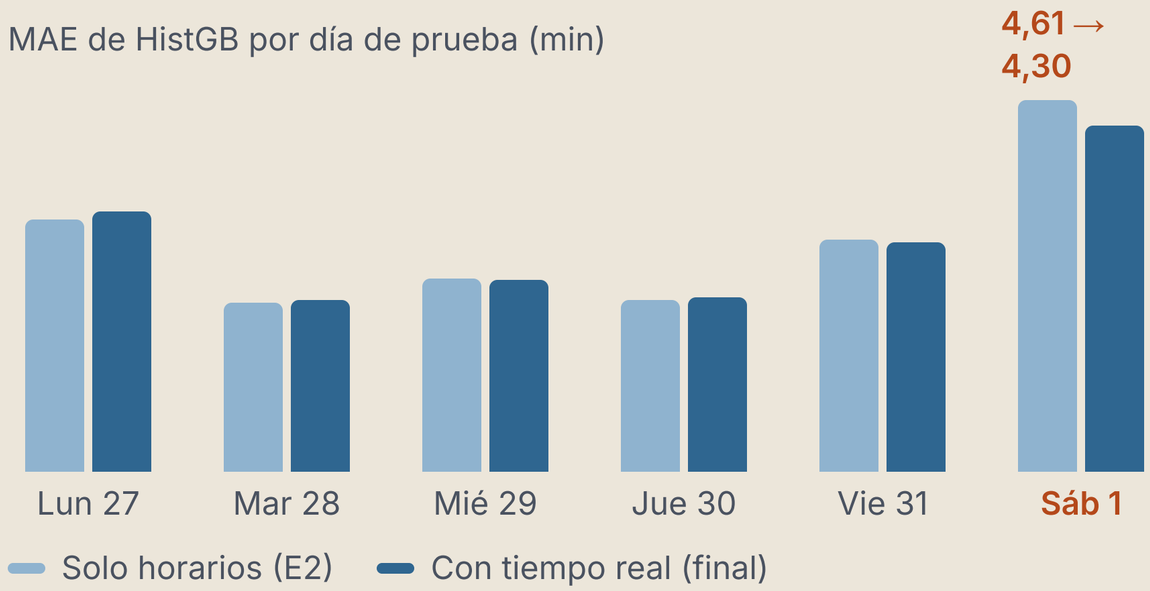
min frente a Ridge + TR: la relación no es lineal

Retraso de buses en Cusco · IC 95 % por bootstrap pareado por día

RESULTADOS · VARIABLES EN TIEMPO REAL

El tiempo real ayuda donde la hora y el día no bastan

MAE de HistGB por día de prueba (min)



2,77 → 2,76

MAE global de HistGB: empate (IC −0,06 a 0,11; 3 de 6 días)

3,41 → 3,22

Ridge sí mejora: no capturaba la interacción hora × día

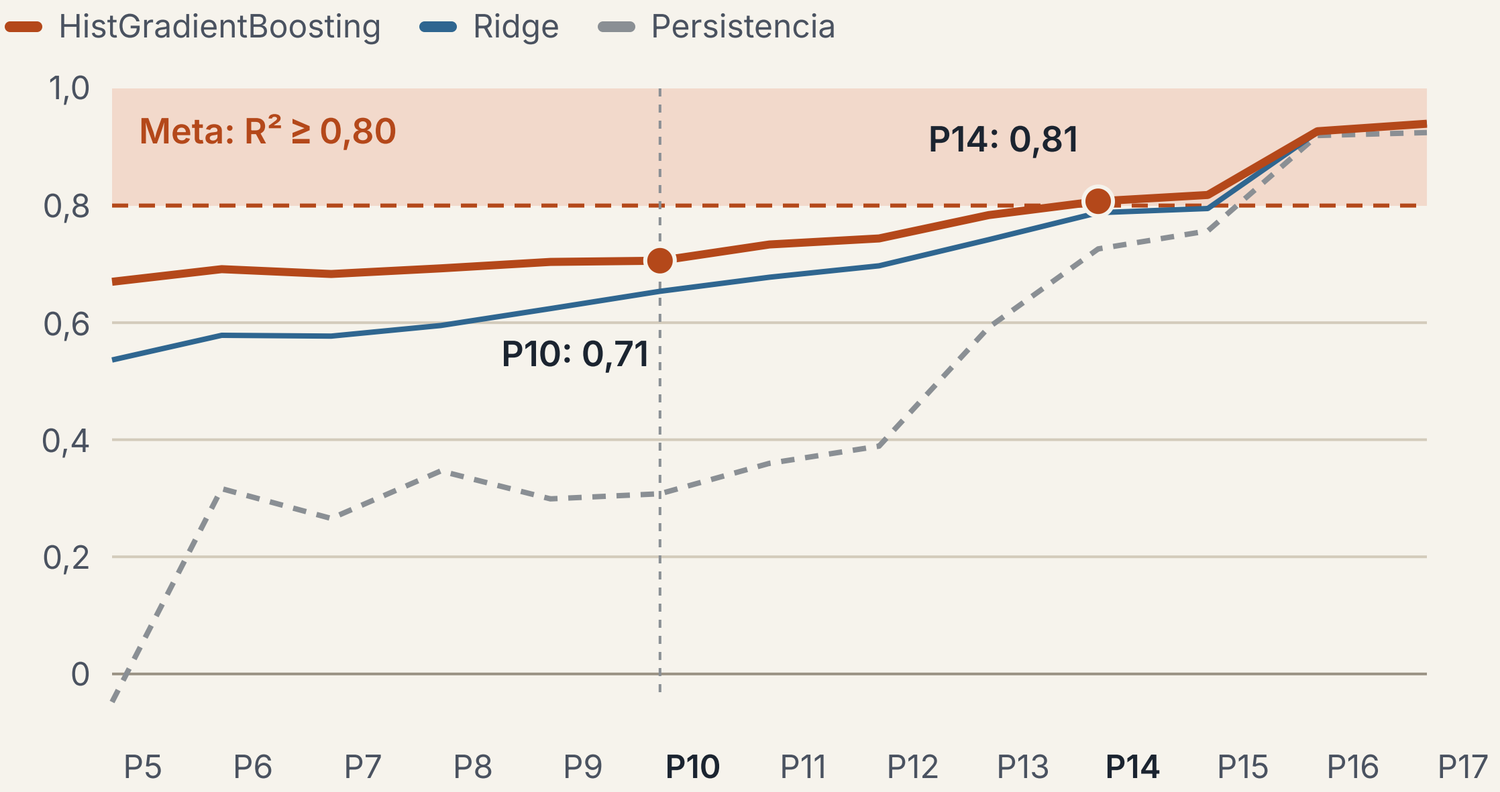
0,74 → 0,41

peso del día de la semana: el modelo usa la congestión observada en vez del calendario

Retraso de buses en Cusco · importancia por permutación en validación cruzada

RESULTADOS · META $R^2 \geq 0{,}80$

Desde P10 no se alcanza; desde P14, sí



Desde P10, ningún pliegue llega a 0,80 (máximo 0,78); en la prueba, 0,76 con IC 0,62–0,84: no se demuestra.

El R^2 sube cuando el bus ya pasó el tramo P12–P14, donde se produce el retraso.

Ajustar los hiperparámetros para cada punto cambia el MAE en 0,07 min o menos.

Retraso de buses en Cusco · R^2 en validación cruzada, punto de partida P5 a P17

Casi nunca confunde una vuelta retrasada con una adelantada

Matriz de confusión en la prueba (filas: clase real)

	Adelantado	Puntual	Retrasado
Adelantado	120 52 %	105 45 %	8 3 %
Puntual	24 5 %	330 75 %	84 19 %
Retrasado	0 0 %	49 20 %	202 80 %

Clases TCRP: adelantado < −1 min · puntual · retrasado > 5 min

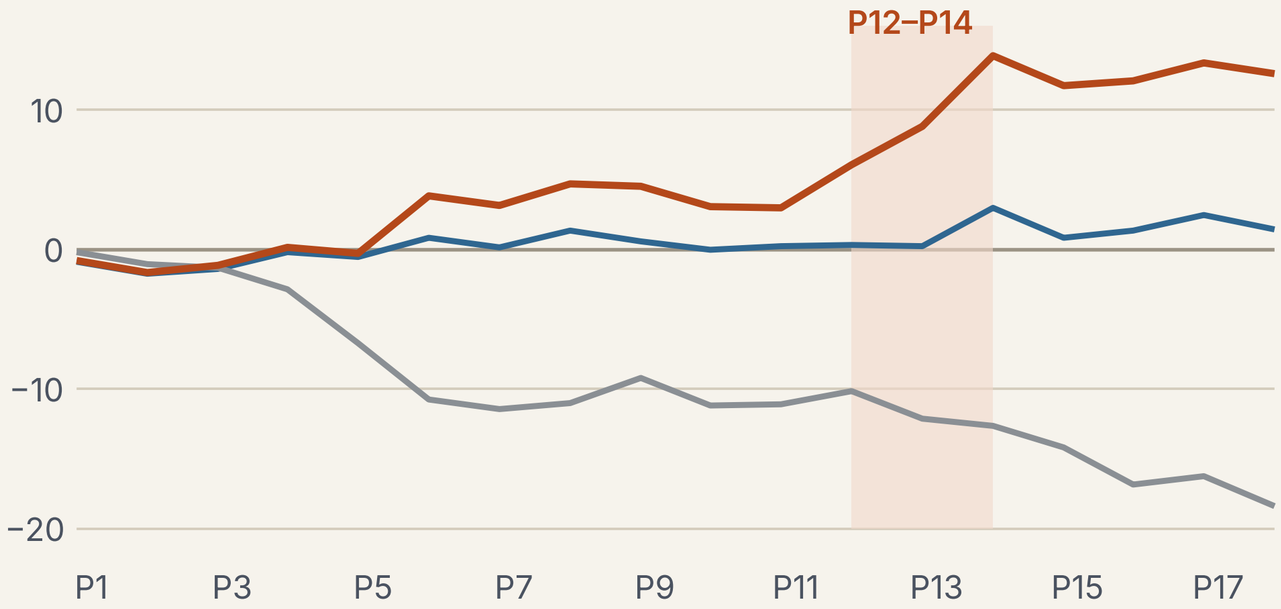
F1 macro en la prueba [IC 95 %]

Modelo	F1 macro
Clase mayoritaria	0,22
Persistencia	0,50
Regresión logística	0,65
Random Forest	0,66
HistGB clasificador	0,69
HistGB regresión + umbrales	0,70

> Elegido en validación cruzada (0,69) y mejor también en la prueba: 0,70 [0,62; 0,77].

Un grupo de retraso alto concentra el error

Retraso medio (min) por punto de control de cada grupo · K-means, k = 3



Retraso alto · 14 %

Salta entre P12 y P14; 72 % son sábados. MAE en la prueba: 4,84 frente a 2,26 min

Adelantado · 7 %

Madrugada, plantilla de 147 min: hasta −18 min en P18

Normal · 79 %

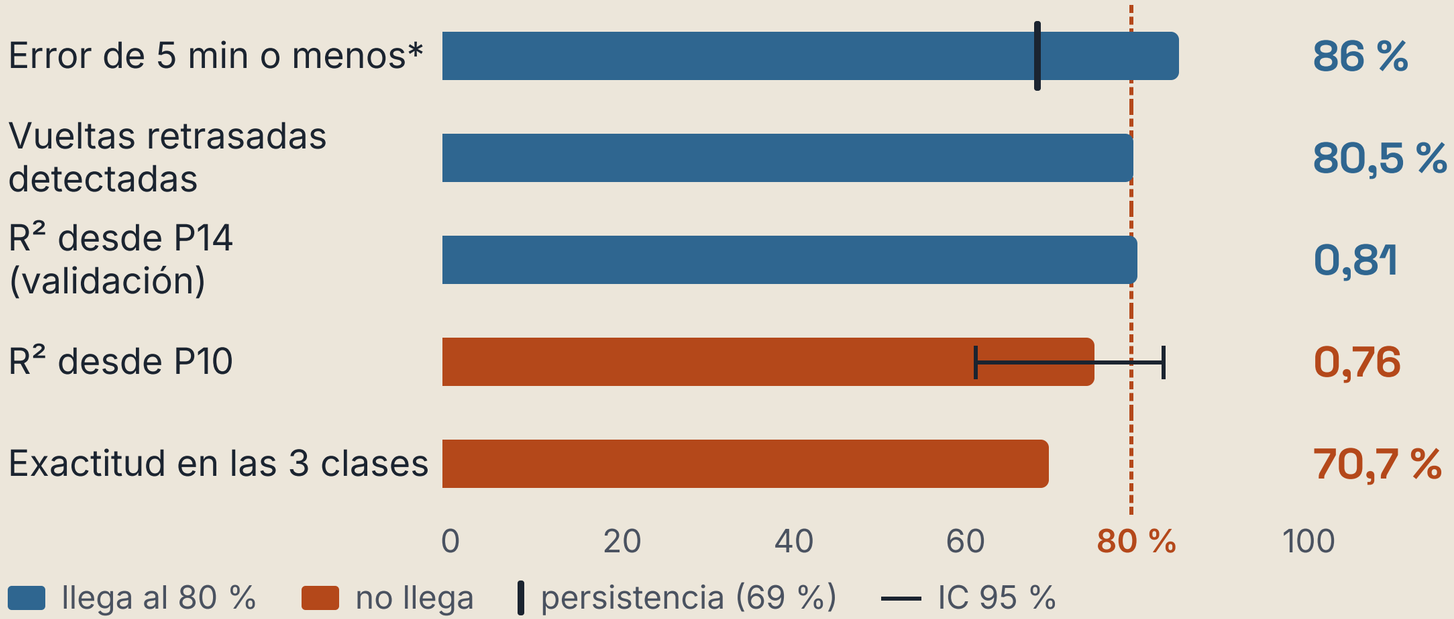
Cerca del horario en toda la ruta

Silueta 0,69 con k = 2 (solo separa los adelantados) y 0,54 con k = 3

Retraso de buses en Cusco · 2 361 perfiles completos del desarrollo; la prueba se asigna al centroide más cercano

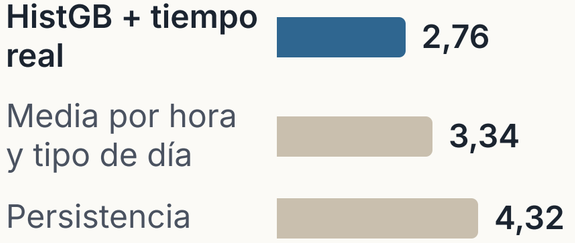
Dos métricas superan el 80 %; R² y exactitud no

Cada métrica frente a la referencia del 80 % (el R² se lee como proporción: 0,80 = 80 %)



Meta 1: superar a la mejor línea base

MAE en la prueba (min, menos es mejor)



Cumplida: 0,58 min menos

Retraso de buses en Cusco · prueba 27 ene–1 feb · *umbral de 5 min de TCRP, calculado después de la evaluación única: descriptivo

DISCUSIÓN

El límite está en la información disponible en P10

Por qué ganan los ensambles

El retraso cambia en la segunda mitad según hora, plantilla y día: interacciones no lineales que la persistencia y Ridge no capturan.

Por qué el tiempo real ayuda poco

Los precedentes salen a una hora parecida el mismo día: su información ya estaba en la hora y el día. Ayuda donde estos no bastan.

El horario también falla

La plantilla de 147 min termina adelantada: parte de lo que se predice es error del horario. Si cambia, hay que reentrenar.

LIMITACIONES

- Una sola ruta y 27 días: sin estacionalidad
- Prueba de 6 días con un solo sábado: IC aproximados
- Horario nuevo de 168 min solo en la prueba
- La prueba ya se usó en el E1 y el diagnóstico del E2 inspiró el tiempo real
- Se excluye el 11 % de vueltas sin P18

Anticipa retrasos grandes con 8 puntos de antelación

2,76 min
de error medio desde P10: 36 % menos que la persistencia

80 %
de las vueltas retrasadas se detectan ya en P10

0,81
de R^2 desde P14; desde P10 es 0,76

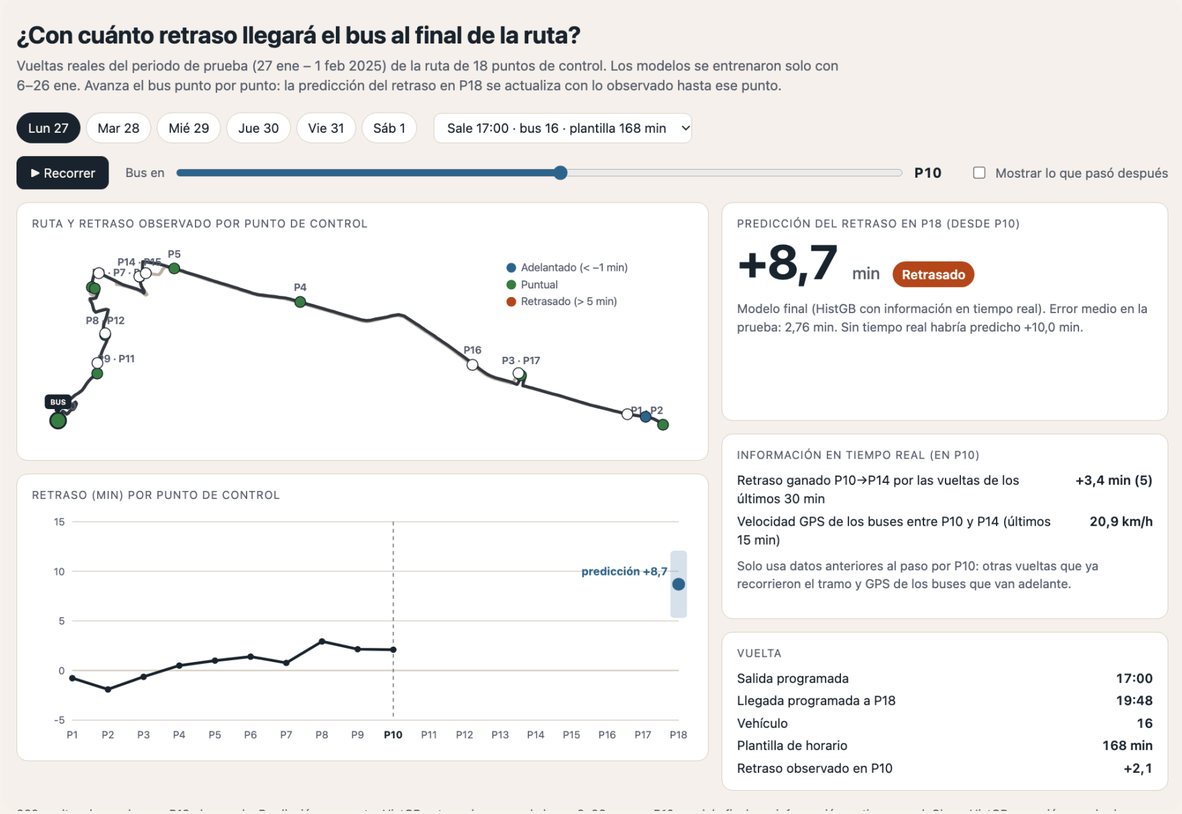
−0,31 min
de error el sábado con tiempo real (4,61 → 4,30)

TRABAJO FUTURO

- Más semanas, sobre todo sábados y feriados
- Horario descrito por tramos
- Predecir todos los puntos y el apelotonamiento
- Validar la demo con despachadores

DEMOSTRACIÓN

Demo: las vueltas reales de la prueba



1. Elegir un día de prueba y una vuelta
2. Avanzar el bus punto por punto
3. Ver la predicción en P18, su error esperado y la clase
4. En P10: el modelo final con información en tiempo real
5. Revelar lo que pasó y compararlo

> Un solo archivo HTML que funciona sin internet: Entrega final/demo/index.html