



ugr

Universidad
de **Granada**

TRABAJO FIN DE GRADO
INGENIERÍA EN INFORMÁTICA

Sistema de Aprendizaje Automático para la Asesoría Personalizada de Dosis de Insulina en Personas con Diabetes Tipo 1

Autor

Carlos Muñoz Sánchez

Directores

Salvador García López



Escuela Técnica Superior de Ingenierías Informática y de
Telecomunicación

—

Granada, Junio de 2025

Agradecimientos

Quisiera agradecer a todas las personas que han estado conmigo durante el desarrollo de este trabajo. A mi tutor por su inestimable ayuda y asesoramiento y a mi familia y amigos por estar siempre al pie del cañón.

Índice

| | |
|---|-----------|
| Agradecimientos..... | 2 |
| Índice..... | 3 |
| Resumen..... | 4 |
| Abstract..... | 5 |
| 1. Introducción..... | 6 |
| 1.1 Contexto sobre la diabetes tipo 1..... | 6 |
| 1.2 Motivación..... | 7 |
| 1.3 Estado del arte..... | 7 |
| 1.4 Propuesta y objetivos..... | 8 |
| 2. Metodología..... | 9 |
| 2.1 Los datos..... | 9 |
| 2.2 Los modelos..... | 14 |
| 2.3 Entrenamiento, evaluación y obtención de predicciones..... | 16 |
| 2.4 Implementación..... | 21 |
| 3. Marco experimental..... | 22 |
| 3.1 Análisis de los datos..... | 22 |
| 3.1.1 Valores nulos..... | 22 |
| 3.1.2 Análisis estadístico..... | 22 |
| 3.2 Preprocesado de datos..... | 25 |
| 3.3 Detección de anomalías..... | 26 |
| 3.4 Establecimiento de una referencia..... | 28 |
| 4. Experimentación..... | 32 |
| 4.1 Experimento 1: Modelos generales con un subconjunto de características..... | 32 |
| 4.1.1 Resultados..... | 32 |
| 4.2 Experimento 2: Modelos generales con todas las características..... | 33 |
| 4.2.1 Resultados..... | 33 |
| 4.2.2 Análisis de la importancia de las características..... | 34 |
| 4.3 Experimento 3: Modelos personalizados..... | 36 |
| 4.3.1 Resultados..... | 37 |
| 4.4 Experimento 4: Modelos generales con discretización..... | 39 |
| 4.4.1 Resultados..... | 41 |
| 5. Análisis de los resultados..... | 44 |
| 6. Conclusiones..... | 45 |
| 7. Referencias..... | 47 |
| 8. Anexo..... | 49 |

Sistema de Aprendizaje Automático para la Asesoría Personalizada de Dosis de Insulina en Personas con Diabetes Tipo 1

Carlos Muñoz Sánchez

Palabras clave: Diabetes Mellitus, aprendizaje automático, predicción de glucosa postprandial, sistema de recomendación de insulina, Ohio T1DM, TabPFN, TabNet, XGBoost, SVM, Random Forest

Resumen

Trasfondo:

La principal consecuencia de la diabetes tipo 1 es la incapacidad del cuerpo para producir insulina, lo que provoca un descontrol en los niveles de glucosa en sangre. Las personas que la padecen necesitan administrarse insulina con herramientas externas para poder mantener sus niveles de azúcar en sangre en rangos saludables. Decidir la dosis exacta de insulina es una tarea que deben hacer a diario, atendiendo a múltiples factores que determinan la cantidad adecuada.

Objetivo:

El objetivo de este estudio es diseñar un sistema capaz de recomendar dosis de insulina a diabéticos de tipo 1 frente a la situación de una ingesta de alimentos.

Métodos:

Utilizando datos de la base de datos Ohio T1DM en su versión de 2018, se han entrenado diferentes algoritmos de aprendizaje automático para que sean capaces de predecir el nivel de glucosa de una persona a las dos horas de una ingesta. Para realizar predicciones, los algoritmos toman como entrada información relativa a la ingesta, al usuario y a su entorno y predicen el nivel de glucosa tras la ingesta para un amplio rango de dosis de insulina, proporcionando como recomendación aquella dosis o rango de dosis que mantienen al usuario en el rango óptimo.

Resultados:

El mejor modelo entrenado obtiene un RMSE de 51,95 mg/dL. Ningún modelo es capaz de realizar predicciones con la suficiente confianza clínica como para poder ser utilizados en situaciones reales. Es difícil comparar el rendimiento de estos modelos con otros del estado del arte, ya que habitualmente las predicciones de insulina no se centran únicamente en situaciones de ingestas, como es el caso de esta investigación. El conjunto de datos utilizado es significativamente menor que los empleados habitualmente en otras investigaciones similares pero, salvando las diferencias, los resultados no parecen ser peores en comparación, demostrando que el refinamiento realizado sobre los datos es de gran valor, particularmente en escenarios como este.

Conclusiones:

Dadas las características de los entrenamientos, se concluye que el principal factor limitante del rendimiento de los modelos es la carencia de información relevante en los datos usados para entrenar. Asimismo, se confirma que la diabetes requiere de un tratamiento personalizado y metódico para que los diabéticos puedan mantener su glucosa en rango, reafirmando así la complejidad del problema y el reto para la salud pública que supone.

Machine Learning System for Personalized Insulin Recommendations for People with Type 1 Diabetes

Carlos Muñoz Sánchez

Keywords: Diabetes Mellitus, machine learning, postprandial glucose prediction, insulin recommender system, Ohio T1DM, TabPFN, TabNet, XGBoost, SVM, Random Forest

Abstract

Background:

The main consequence of type 1 diabetes is the body's inability to produce insulin, leading to poor regulation of blood glucose levels. Individuals that suffer from this condition must administer insulin through insulin delivery devices to maintain blood sugar levels within healthy ranges. Determining the exact insulin dose is a daily task that requires considering multiple factors.

Objective:

This study aims to develop a system capable of recommending insulin doses for individuals with type 1 diabetes in response to food intake.

Methods:

Using data from the 2018 version of the Ohio T1DM dataset, several machine learning algorithms were trained to predict the individual's blood glucose levels two hours after a meal. To generate predictions, the algorithms take as input variables related to the meal, the user, and their context, and estimate postprandial glucose levels across a wide range of insulin doses. The system then recommends the dose or range of doses expected to maintain glucose levels within the optimal range.

Results:

The best trained model obtained an RMSE of 51.95 mg/dL. No model demonstrated sufficient predictive accuracy or clinical reliability to be used in real-world settings. It is difficult to compare the performance of these models with others in the state of the art, as insulin predictions are not usually focused solely on intake situations, as is the case in this research. The dataset used is significantly smaller than those commonly employed in other similar investigations but, differences aside, the results appear to be no worse in comparison, demonstrating that the refinement performed on the data is of great value, particularly in scenarios such as this one.

Conclusions:

Based on the training characteristics, the main performance limitation appears to be the lack of relevant information in the training data. The findings also reaffirm that diabetes management demands a highly personalized and methodical approach to maintain blood glucose within target ranges, underscoring both the complexity of the problem and its significance as a public health challenge.

1. Introducción

1.1 Contexto sobre la diabetes tipo 1

Las personas con T1DM (*Type 1 Diabetes Mellitus, Diabetes de tipo 1*) presentan un problema autoinmune por el cual su sistema inmunológico ataca a las células beta del páncreas, que son las encargadas de producir insulina [1]. Esto implica una pérdida de la capacidad para metabolizar el azúcar presente en la sangre, lo que obliga a los pacientes a tener que regular sus niveles de glucosa con herramientas externas que les suministren la insulina que necesitan.

Habitualmente, la administración de insulina se realiza mediante una de las siguientes técnicas:

1. Jeringas o bolígrafos de insulina: Se aplican directamente en el abdomen, muslos, glúteos o brazos y suministran insulina mediante una inyección.
2. Bombas de insulina: Dispositivos capaces de administrar insulina de forma continua mediante un catéter insertado bajo la piel. Los modelos más avanzados incluyen herramientas de monitorización de glucosa para un mayor control.

Hay diferentes tipos de insulina que se diferencian por el tiempo que tardan en empezar a hacer efecto y la duración del mismo. Normalmente, las personas con T1DM combinan insulina de acción rápida y de acción prolongada:

1. **Insulinas de acción prolongada:** Proporcionan un nivel estable de insulina durante horas, por lo que se utiliza para imitar la secreción basal del páncreas. Los pacientes en tratamiento con bolígrafos de insulina suelen suministrarse una vez al día una cantidad calculada con un experto sanitario, aunque esta cantidad puede variar a lo largo de la vida del paciente. Las personas en tratamiento con bombas de insulina suelen recibirla de forma constante a lo largo del día. Muchas bombas también pueden realizar un suministro inteligente en función de las tendencias de glucosa en sangre (por ejemplo, para detener el suministro durante episodios de hipoglucemia).
2. **Insulinas de acción rápida:** Empiezan a actuar a los pocos minutos de la inyección y su efecto se prolonga unas pocas horas. Se utilizan principalmente para controlar los picos de glucemia producidos por la ingesta de alimentos y también para corregir estados de hiperglucemia. La dosis adecuada de este tipo de insulina depende de muchos factores: carbohidratos de los alimentos, tendencia actual de glucosa, sensibilidad a la insulina, actividad física prevista, momento y tipo de ingesta, etc.

El manejo adecuado de la insulina permite a las personas con T1DM llevar una vida normal y saludable. La mayor dificultad a la que suelen enfrentarse reside en el cálculo de la dosis de insulina rápida. Esto se debe a que la dosis adecuada depende de una serie de factores no solo asociados a lo que se ingiere, sino también a distintas circunstancias personales y ambientales. Esto hace que los pacientes acaben calculando sus dosis de insulina en base a reglas empíricas que van aprendiendo en su experiencia con la enfermedad, las cuales no

siempre son eficaces para capturar la complejidad subyacente al metabolismo de la glucosa.

1.2 Motivación

La T1DM puede ser una enfermedad incapacitante. Las fluctuaciones en los niveles de azúcar en sangre pueden causar episodios de hipoglucemia o hiperglucemia, que pueden afectar la concentración, la energía y la capacidad para trabajar o estudiar. Además, un descontrol prolongado de los niveles de glucosa suele traer consigo complicaciones a largo plazo, como problemas cardíacos, renales, de visión y nerviosos.

Es aquí donde nace la necesidad de diseñar herramientas que puedan ayudar a personas con T1DM a mantener sus niveles de glucosa en un rango saludable.

1.3 Estado del arte

El trabajo “*Insulin Recommender Systems for T1DM: A Review Joaquim Massana, Ferran Torrent-Fontbona, and Beatriz López*” , publicado en 2020 por investigadores de la Universidad de Girona [2] analiza 70 trabajos acerca de sistemas de recomendación de insulina y discute el estado del arte en esta materia. En este trabajo, se observa que la mayoría de sistemas existentes son sistemas expertos y sistemas basados en reglas, aunque la comparación entre estos y otras propuestas es muy complicada debido a una falta de estándares dentro de la investigación, siendo habitual el uso de medidas de rendimiento distintas y metodologías diversas (y con razonamientos demasiado opacos en muchos casos).

Por otro lado, otras investigaciones se centran en la predicción de niveles de glucosa en lugar de en la recomendación de dosis de insulina, aunque, como se explicará en secciones posteriores, este tipo de sistemas también pueden utilizarse para el asesoramiento de dosis de insulina. En esta línea, es más habitual ver soluciones basadas en aprendizaje automático, sobre todo las más modernas [3].

La delicada naturaleza de los datos relacionados con la diabetes es una traba a la hora de ser compartidos entre diferentes estudios, lo que compromete la calidad de las investigaciones y dificulta la comparación entre diferentes propuestas [4]. Este hecho, sumado a la dificultad inherente de un problema en el que intervienen procesos tan complejos y variables como el metabolismo en los seres humanos provoca que las soluciones disponibles sean todavía inviables para el uso real. Tal y como se apunta en el trabajo “*GLYFE: review and benchmark of personalized glucose predictive models in type 1 diabetes, M. De Bois, M. A. E. Yacoubi, and M. Ammi*”, el RMSE obtenido por los modelos para un horizonte temporal de predicción de 2h (como el de este trabajo) oscila entre el 46,24 y el 47,56. Aunque puede llegar a resultar útil como una segunda opinión, estos resultados aún están lejos de tener suficiente confianza clínica, por lo que este problema no puede considerarse resuelto con las propuestas actuales del estado del arte.

1.4 Propuesta y objetivos

Tal y como se ha descrito en la sección 1.2, la T1DM puede ser una enfermedad de peligrosas consecuencias si no se lleva a cabo un tratamiento adecuado, pudiendo causar ya no sólo incomodidades en la vida cotidiana, sino secuelas a largo y corto plazo e incluso la muerte.

Es aquí donde nace la necesidad de diseñar herramientas que puedan ayudar a personas con T1DM a mantener sus niveles de glucosa en un rango saludable.

El objetivo original de este trabajo consiste en desarrollar un sistema de aprendizaje automático capaz de ofrecer recomendaciones personalizadas de insulina a personas con T1DM frente a ingestas, que es uno de los escenarios más complejos a los que se enfrentan cada día y que más impacto puede tener en su salud.

Habitualmente, las investigaciones relacionadas con la predicción y control de glucosa no se centran específicamente en las ingestas. Con el objetivo de tratar este punto concreto, se ha llevado a cabo un minucioso refinamiento de los datos disponibles para emplear información que habitualmente no se tiene en cuenta a pesar de su importancia, como puede ser el tiempo de espera. El tiempo de espera es la diferencia temporal entre el momento de la ingesta y el momento de la inyección de insulina. Un buen tiempo de espera favorece que el pico de glucosa de la ingesta coincida con el pico de acción de la insulina, mejorando la regulación de los niveles de azúcar.

2. Metodología

2.1 Los datos

Este trabajo se desarrollará sobre la base de datos OhioT1DM [5], la cual recoge 8 semanas de registros de 6 personas en tratamiento con bombas de insulina y que contaban con bandas de salud Basis Peak. Dicha información se presenta en archivos en formato XML donde cada paciente tiene uno con los datos de entrenamiento y otro con los datos para test. Cada archivo está identificado con el identificador único de cada paciente y contiene registros con la siguiente información y una marca temporal asociada:

1. Niveles de glucosa, cada 5 minutos. Están disponibles los niveles registrados por el CGM (Monitor Continuo de Glucosa, en inglés) y por el paciente a través de muestras de sangre.
2. Niveles de insulina basal base y temporal durante todo el periodo de monitorización.
3. Inyecciones de insulina de acción rápida asociadas a ingestas. También se incluye el tipo de inyección, ya que las bombas de insulina pueden suministrar las dosis de forma prolongada en el tiempo o de una sola vez.
4. Registro manual de ingestas por parte de los pacientes, con una estimación de carbohidratos.
5. Registro manual de tiempo y calidad del sueño.
6. Registro manual de jornadas laborales y una valoración del esfuerzo físico asociado.
7. Registros de episodios de estrés.
8. Episodios de hipoglucemias.
9. Registros de episodios de enfermedad.
10. Sesiones de ejercicio con duración e intensidad.
11. Pulso, registrado cada 5 minutos
12. Respuesta galvánica de la piel, cada 5 minutos
13. Temperatura capilar, cada 5 minutos.
14. Temperatura ambiente, cada 5 minutos.
15. Pasos dados cada 5 minutos.
16. Tiempo y calidad de sueño mediante registro automático por la banda de salud.

El objetivo de este trabajo es obtener un modelo que sea capaz de, frente a una ingesta, recomendar una dosis de insulina y tiempo de espera para asegurar la normoglucemia tras esta. El formato más conveniente que pueden tener los ejemplos de entrenamiento para un modelo así sería aquel en el que cada ejemplo represente una ingesta con toda su información asociada y una etiqueta que diga cuál es el valor adecuado de insulina para ese caso y otra con el tiempo de espera.

La base de datos OhioT1DM contiene prácticamente toda la información necesaria para construir ejemplos de este estilo, pero con una característica muy importante: las inyecciones de insulina que cada paciente ha escogido para hacer frente a las ingestas no siempre son las adecuadas. Y es que, si los pacientes ya supieran cuál es la dosis precisa, no habría ningún problema que solucionar.

Entrenar modelos directamente bajo estas circunstancias daría como resultado modelos que, en el mejor de los casos, predicen tan bien como los individuos, pues su entrenamiento consistiría en imitar sus decisiones. Una solución a este obstáculo podría ser eliminar todos los ejemplos en los que la dosis de insulina rápida y tiempo de espera no consiga mantener la normoglucemia. Sin embargo, esto sería una mala idea por dos motivos:

1. Se reduciría significativamente la cantidad de información disponible para el entrenamiento.
2. Los modelos podrían ser muy precisos para los casos en los que los pacientes ya hacen buenas predicciones, pero no habría garantías de que las predicciones sean buenas precisamente en esos otros casos más difíciles, que son realmente los que mayor interés práctico tienen.

Por tanto, la solución que este trabajo propone consiste en construir ejemplos en los que la etiqueta no sea la dosis de insulina inyectada, sino el nivel de glucosa tras la ingesta. De este modo, el modelo sería capaz de predecir, dada una ingesta (con una dosis de insulina y tiempo de espera elegidos por el paciente), si los niveles de glucosa se mantendrán o no en un rango aceptable tras la ingesta.

Con un modelo así, se podría partir de un rango de valores posibles de dosis y tiempo de espera y usar el modelo para predecir qué combinación es la idónea. El resultado sería un rango de valores en los que el modelo estima que la glucosa se mantendrá en un nivel adecuado.

La construcción de los ejemplos de entrenamiento y su etiquetado es entonces la primera tarea de este trabajo, la cual se ha llevado a cabo mediante un *script* que lee los datos en crudo de los archivos XML y los transforma para obtener datos etiquetados en formato tabular con la siguiente estructura:

| Nombre de la característica | Tipo de dato | Método de obtención | Observaciones |
|--------------------------------------|--------------|--|--|
| Id | Entero | Indicado directamente en el archivo XML de cada paciente | No se utiliza para entrenar, sino para poder dividir los datos por paciente cuando sea necesario |
| Media de pasos cada 5 minutos | Decimal | Obtenido a través del registro de pasos dados cada 5 minutos | Se decide usar la media de pasos en lugar de la cantidad de pasos cercanos a una ingesta porque no es habitual que una persona sepa el nº de pasos que va a dar en un intervalo de tiempo concreto, mientras que el nº de pasos aproximado que una persona da al día es un valor mucho más accesible |
| Nivel de | Entero | Se usa el registro | Todas las medidas de glucosa |

| | | | |
|---|-----------|---|--|
| glucosa en el momento de la ingesta | | más cercano temporalmente a la ingesta | vienen dadas en mg/dL, que es la unidad estándar junto con los mmol/L |
| Tendencia de glucosa | Entero | Diferencia del nivel de glucosa entre la glucosa asociada a la ingesta y el registro realizado 30 minutos antes de esta | Se emplea una ventana de 30 minutos porque el intervalo habitual en herramientas de monitorización de glucosa (como Dexcom G6 o Freestyle libre) es en torno a esta cantidad |
| Carbohidratos de la ingesta | Entero | Indicado en el registro de cada ingesta | La cantidad es estimada e introducida manualmente por cada paciente |
| Tipo de ingesta | Enumerado | Indicado en el registro de cada ingesta | Puede ser desayuno, snack, comida o cena |
| Insulina basal en el momento de la ingesta | Decimal | Se suma la cantidad de insulina basal y basal temporal activas en el momento de la ingesta | |
| Horas de sueño el día de la ingesta | Decimal | Se obtiene de los eventos de sueño registrados por el paciente | |
| Calidad de sueño el día de la ingesta | Entero | Se obtiene de los eventos de sueño registrados por el paciente | Puede ser: 1. Poco descanso 2. Descanso aceptable 3. Buen descanso |
| Cantidad de horas de trabajo el día de la ingesta | Decimal | Se obtiene de los eventos de trabajo registrados por el paciente | |
| Intensidad física del trabajo el día de la ingesta | Entero | Se obtiene de los eventos de trabajo registrados por el paciente | Toma un valor en una escala del 1 al 10, siendo 10 la mayor actividad física posible |
| Eventos de estrés el día de la ingesta | Binario | Se obtiene de los eventos de estrés registrados por el paciente | |
| Enfermedad del paciente en el momento | Binario | Se obtiene de los eventos de enfermedad | |

| | | | |
|---|---------|--|---|
| de la ingesta | | registrados por el paciente | |
| Duración de ejercicio físico | Entero | Se obtiene de los eventos de ejercicio registrados por el paciente | Duración representada en minutos |
| Intensidad del ejercicio físico | Entero | Se obtiene de los eventos de ejercicio registrados por el paciente | Toma un valor en una escala del 1 al 10, siendo 10 la mayor intensidad física posible |
| Pulsaciones por minuto en el momento de la ingesta | Entero | Indicado en el registro de pulsaciones por minuto | |
| Respuesta galvánica de la piel en el momento de la ingesta | Decimal | Indicado en el registro de GSR | Medida en microsiemens |
| Temperatura corporal en el momento de la ingesta | Decimal | Indicado en el registro de temperatura corporal | En grados Fahrenheit |
| Temperatura ambiente en el momento de la ingesta | Decimal | Indicado en el registro de temperatura ambiente | En grados Fahrenheit |
| Insulina asociada a la ingesta | Decimal | Asociado al registro de la ingesta | Una de las variables que se pretende dar como parte del asesoramiento |
| Tiempo de espera | Decimal | Diferencia temporal entre la ingesta y la inyección de insulina | Una de las variables que se pretende dar como parte del asesoramiento |

Tabla 1: desglose de las características extraídas del dataset Ohio T1DM

La etiqueta asociada a cada ejemplo es indicadora del nivel de glucemia del usuario a las 2 horas de la ingesta, que es el intervalo temporal recomendado por la mayoría de profesionales sanitarios para el seguimiento de los valores de glucosa.

Se han valorado las siguientes formas de realizar el etiquetado:

1. **Etiquetado “en crudo”:** La etiqueta es directamente el nivel de glucosa a las 2 horas de la ingesta.

2. **Etiquetado “por distancia”:** La etiqueta vale 0 si el paciente se encuentra en rango a las 2 horas de la ingesta. Si no se encuentra en rango, se etiqueta con la distancia del nivel de glucosa al umbral de normoglucemia (70-180), de forma que esta etiqueta puede tomar valores negativos.
3. **Etiquetado “binario”:** Se marcan como positivos los ejemplos en los que el paciente se mantiene en rango y como negativos aquellos en los que no.

Cada una de estas etiquetas ofrece ventajas y desventajas. Por ejemplo, el etiquetado binario puede simplificar el aprendizaje al haber sólo dos clases, mientras que los otros ofrecen información más detallada sobre la glucosa.

Finalmente, la decisión ha sido utilizar únicamente el etiquetado “en crudo”.

El objetivo final de los modelos entrenados es recomendar dosis de insulina. Dadas las características de los datos con los que los modelos son entrenados, estas recomendaciones se hacen de forma indirecta. En lugar de obtener directamente el valor de insulina recomendado, se debe realizar una simulación para cada ingesta en el que el modelo realiza una predicción con diferentes valores posibles de insulina y tiempo de espera. Tras ello, se obtiene un cierto rango de valores para los que el modelo estima que el usuario mantendrá niveles estables de glucosa.

Obtener un valor único de insulina, que sería al fin y al cabo el propósito final del modelo, es un paso extra que debe ser pensado detenidamente. Por ejemplo, un modelo podría predecir que dosis entre 2 y 10 unidades de insulina mantienen al paciente en rango. Dentro de este rango, ¿qué valor se debe dar como recomendación final? Una idea sería un valor central, como 6. Pero, en realidad, quizás lo más sensato sea dar aquel valor que tenga más margen de error. Por ejemplo, un modelo puede calcular que 6 unidades de insulina mantienen al paciente en rango, pero no es lo mismo que se mantenga a 176 de glucosa que a 120. En el segundo caso, un pequeño error en la precisión de los cálculos tiene mucho margen tanto por encima como por debajo, ya que valores entre 70 y 180 se consideran adecuados tras una ingesta. En el primer caso, un pequeño error puede provocar una ligera hiperglucemia en el paciente. Para realizar este tipo de cálculos, es necesario que la predicción del modelo sea suficientemente interpretable.

Los modelos entrenados con el etiquetado binario no ofrecen esta posibilidad, puesto que su respuesta es simplemente “sí” o “no”. No hay manera de saber cuánto margen de error tienen las predicciones. El porcentaje de confianza con el que el modelo clasifica el ejemplo en la clase positiva o negativa no necesariamente habla del valor de glucosa asociado a ese ejemplo concreto. Los modelos entrenados con el etiquetado por distancia tienen un problema semejante. La forma en la que predicen que un paciente queda en rango es respondiendo con un valor muy cercano a 0. Pero, de nuevo, ese 0 significa que el paciente acaba en normoglucemia, pero no hay forma de saber con qué margen de error. Además, los modelos de este tipo difícilmente responden con un 0,0 exacto, lo que supone otro problema: ¿Cuánta cercanía al 0,0 se considera un caso en el que el paciente acaba en situación de normoglucemia? ¿1.5 se considera normoglucémico? ¿Y -3.7?

Es por esto por lo que usar el etiquetado en crudo es más práctico, ya que se obtiene directamente la predicción sobre el nivel de glucosa de los pacientes tras la ingesta. Con

esta información, se puede sugerir al paciente el valor de insulina con el que el modelo estima que quedaría en un valor de glucosa con mucho margen de error, como 120. Además, en caso de tener suficiente confianza en el modelo, se podría incluso ofrecer la posibilidad de que el modelo recomiende un valor de insulina con el que el paciente queda en un valor concreto de glucosa. Por ejemplo, un paciente podría querer reducir el riesgo de hipoglucemia porque va a estar en un contexto en el que le es difícil o incómodo comer (en una excursión, trabajando, etc). En ese caso, podría pedirle al modelo la cantidad de insulina con la que se quede a 160 puntos de glucosa.

No obstante, tal y como se explicará más adelante, los modelos del experimento 4 han sido entrenados con un enfoque parecido al binario. En un intento de simplificación del problema para mejorar los resultados, estos experimentos utilizan un etiquetado en el que los valores en crudo de glucosa se han discretizado para obtener entre 5 clases que representan distintos rangos de glucosa. Estos modelos tratan de mejorar la confianza de sus predicciones a cambio de ofrecer recomendaciones menos específicas, puesto que recomiendan rangos de insulina en lugar de un valor concreto.

2.2 Los modelos

La decisión acerca de qué modelos utilizar es crucial para el trabajo. Existen numerosas opciones, cada una con sus ventajas y desventajas, características, peculiaridades y ámbitos de aplicación. Para esta investigación, se ha realizado una selección de modelos que, en la medida de lo posible, cumplen con los siguientes requisitos:

1. **Código abierto:** Puesto que la implementación de los modelos no es el objeto de este trabajo, es necesario que su código y documentación estén publicados en internet.
2. **Su entrenamiento no requiere de infraestructura compleja ni se prolonga demasiado en el tiempo:** Todo el código de este trabajo ha sido ejecutado en Google Colab [6] haciendo uso de las unidades de computación que este servicio ofrece de forma gratuita. Entrenar modelos demasiado grandes consumiría la cuota gratuita, bloqueando así el resto del desarrollo.
3. **Trabajan bien con pocos datos:** Dado que el conjunto de entrenamiento usado no supera los 2000 ejemplos.
4. **Admiten el formato de los datos:** Es necesario que los modelos acepten datos tabulares, pues esa es la estructura de los datos empleados.

Además, también se ha valorado que los modelos ofrezcan interpretabilidad y la posibilidad de graficar sus curvas de entrenamiento. Estas dos características permiten valorar con mayor rigor los resultados de los modelos para tomar decisiones informadas acerca del rumbo del trabajo. También es importante tratar de elaborar una selección variada donde se puedan comparar modelos que funcionen con enfoques distintos, desde redes neuronales hasta SVMs.

Bajo estas premisas, los modelos seleccionados son:

- **TabPFN [7]:** Propone una arquitectura de red neuronal diseñada para realizar tareas de clasificación y regresión en conjuntos de datos tabulares. Este modelo utiliza una

red neuronal entrenada previamente con millones de tareas sintéticas generadas a partir de un prior bayesiano, permitiendo realizar inferencia casi instantánea sobre nuevos datos tabulares. Una de las principales ventajas de este modelo es que no requiere de prácticamente ningún ajuste de hiperparámetros, lo cual evita tener que recurrir a técnicas de búsqueda de parámetros como grid-search, que pueden llegar a ser muy costosas computacionalmente. Es una de las propuestas más recientes e innovadoras y ha demostrado ofrecer resultados tan competitivos como otros modelos del estado del arte con un coste computacional cientos o miles de veces menor. Existe un autoensemble denominado AutoTabPFN que suele ser ligeramente más preciso que la versión estándar. Sin embargo, no se ha utilizado en este proyecto por dos motivos: requiere de más esfuerzo computacional y la implementación nativa no incluye herramientas para graficar fácilmente las curvas de entrenamiento. Se ha utilizado la implementación publicada por los propios autores en este repositorio: <https://github.com/PriorLabs/TabPFN>

- **TabNet** [8]: Junto con TabPFN conforman la dupla de redes neuronales empleadas en este trabajo. TabNet tiene una arquitectura basada en redes neuronales profundas cuya principal característica es el uso de un mecanismo de atención secuencial que permite seleccionar de forma dinámica las características más relevantes en cada decisión, lo cual es prometedor en el contexto del dataset OhioT1DM, en el que hay un amplio catálogo de características para cada ejemplo. Se ha utilizado la implementación publicada en este repositorio: <https://github.com/dreamquark-ai/tabnet>
- **SVM** [9]: Los SVM son una opción clásica para problemas de aprendizaje automático. Es una opción especialmente eficaz en espacios de alta dimensionalidad con datasets reducidos, pudiendo capturar patrones complejos mediante el uso de kernels no lineales. Su funcionamiento se basa en encontrar el hiperplano óptimo que separa las clases maximizando el margen entre ellas. Se ha utilizado la implementación de la librería de aprendizaje automático SciKit Learn [10] : <https://scikit-learn.org/stable/modules/svm.html>
- **XGBoost** [11]: Los Gradient-Boosted Trees como XGBoost han sido los modelos más dominantes en tareas tabulares de los últimos años (R. Shwartz-Ziv and A. Armon. Tabular data: Deep learning is not all you need), por lo que su inclusión en este experimento es prácticamente obligatoria. XGBoost construye modelos predictivos a partir de una secuencia de árboles de decisión entrenados de forma iterativa. Además, proporciona herramientas para evaluar la importancia de las características. Se ha utilizado la implementación publicada en el siguiente repositorio: <https://github.com/dmlc/xgboost>
- **RandomForest** [12]: Algoritmo de aprendizaje automático basado en el ensamblado de múltiples árboles de decisión entrenados sobre diferentes subconjuntos del conjunto de datos y de las características. Su fortaleza radica en su capacidad para manejar datos con ruido, capturar relaciones no lineales y gestionar datos con una estructura compleja, características que pueden resultar de utilidad para este problema. Además, ofrece medidas de importancia de variables que pueden contribuir a interpretar los factores más influyentes en las predicciones. Se ha utilizado la implementación de la librería de aprendizaje automático SciKit Learn: <https://scikit-learn.org/stable/modules/ensemble.html#random-forests>

2.3 Entrenamiento, evaluación y obtención de predicciones

Para cada experimento, los modelos han sido entrenados siguiendo una serie de pasos para tratar de garantizar la optimalidad de los entrenamientos teniendo en consideración las limitaciones computacionales y temporales. Estos pasos, por orden de aplicación, son explicados a continuación.

Realizar GridSearch usando 5-fold cross validation. GridSearch es una técnica de optimización de modelos de aprendizaje automático que consiste en entrenar distintas versiones de los modelos empleando diferentes combinaciones de hiperparámetros para después escoger la configuración que mejores resultados ha obtenido. Para comparar el rendimiento de cada versión, se utiliza *5-fold cross validation*, que es una técnica de evaluación estadística utilizada para estimar la capacidad de generalización de un modelo de aprendizaje automático. Consiste en dividir el conjunto de datos disponible en cinco particiones (o "folds") del mismo tamaño, utilizando sucesivamente cuatro de ellas para entrenar el modelo y la quinta para validar. Este proceso se repite cinco veces, de forma que cada partición actúa como conjunto de validación una única vez, y como parte del conjunto de entrenamiento en las restantes. Finalmente, se calculan las métricas de rendimiento promedio, lo que proporciona una robusta y menos sesgada evaluación del comportamiento del modelo frente a nuevas muestras.

Visualizar el entrenamiento del modelo con la mejor configuración encontrada. Tras haber obtenido la mejor configuración, se entrena el modelo con dicha configuración mientras se grafica la evolución de alguna métrica de rendimiento durante el proceso. Esta gráfica es crucial para entender si los modelos están aprendiendo correctamente.

Realizar predicciones. Se obtienen las predicciones del modelo para los conjuntos de entrenamiento y validación.

Evaluar las predicciones. Las métricas de rendimiento utilizadas para evaluar y comparar el rendimiento de los modelos son:

1. Para regresión:

- a. MAE (*Mean Absolute Error*): Es una medida del error medio absoluto del modelo. Se calcula midiendo la distancia entre el valor de glucosa predicho y el real y efectuando el promedio de todas las predicciones. Permite tener una idea de cómo de precisas son las predicciones del modelo en promedio, por lo que será útil para comparar y evaluar los modelos.

$$\text{MAE} = \frac{\sum_{i=1}^n |y_i - x_i|}{n}$$

y = predicción

x = valor real

n = número de instancias

- b. RMSE (*Root Mean Squared Error*): Se calcula tomando el cuadrado de la distancia entre el valor de glucosa predicho y el real. Tras ello, se efectúa el promedio de todas las predicciones y se toma la raíz cuadrada. Al haber

calculado el error mediante el cuadrado de las diferencias en las predicciones antes de promediarlas, se penaliza la existencia de errores grandes, por lo que esta medida ayuda a entender de qué forma falla el modelo. Además, es una de las más utilizadas en otras investigaciones, por lo que permite la comparación con otros modelos del estado del arte.

$$RMSE = \sqrt{\sum_{i=1}^n \frac{(\hat{y}_i - y_i)^2}{n}}$$

y = predicción

x = valor real

n = número de instancias

1. Para clasificación:

- a. Exactitud (*accuracy*): Indica la proporción de predicciones en las que el modelo ha etiquetado correctamente. Puede ser engañosamente alta si las clases están desbalanceadas. Por ejemplo, en un problema con 2 clases con distribución 90 % - 10 %, un modelo que siempre prediga la primera clase estaría obteniendo un 90 % de exactitud. Tal y como se comentará en secciones posteriores, la recomendación de dosis de insulina mediante la aproximación a un problema de clasificación ha sido abordada con clases equilibradas, por lo que esta medida de rendimiento es un buen indicativo de las capacidades predictivas de los modelos.

$$\text{Accuracy} = \frac{\text{Número de Predicciones correctas}}{\text{Número de instancias}}$$

- b. Matriz de confusión: Se trata de una tabla que compara las predicciones del modelo con las etiquetas reales, mostrando cuántas veces se ha acertado o errado en cada clase. Cada fila representa las instancias reales y cada columna las predichas (o viceversa, según convención), permitiendo identificar no sólo la precisión global, sino también qué tipos de errores se cometen con más frecuencia.

| Matriz de confusión | | Etiquetas reales | | | |
|---------------------|------------|------------------|------------|-----|------------|
| | | Etiqueta 1 | Etiqueta 2 | ... | Etiqueta N |
| Etiquetas predichas | Etiqueta 1 | Aciertos | Fallos | ... | Fallos |
| | Etiqueta 2 | Fallos | Aciertos | ... | Fallos |
| | ... | ... | ... | ... | ... |
| | Etiqueta N | Fallos | Fallos | ... | Aciertos |

Adicionalmente al uso de métricas estadísticas clásicas, algunas investigaciones relacionadas con la predicción de glucosa utilizan análisis de rejilla como el Clarke Error Grid Analysis [13]. El Clarke Error Grid tiene en cuenta el impacto clínico de los errores de predicción, clasificando cada predicción en una de cinco zonas (A a E) en función de su potencial para inducir decisiones terapéuticas correctas, benignas o peligrosas.

Las zonas se definen sobre un diagrama cartesiano donde el eje X representa el valor real de glucosa y el eje Y la predicción. La zona A representa predicciones de glucosa con una desviación menor al 20% o que representan hipoglucemia cuando el valor real también se encuentra en ese rango. La zona B agrupa predicciones que, a pesar de presentar un error de más del 20%, no inducirían a una intervención clínica inapropiada. Las zonas C, D y E, en cambio, representan errores con consecuencias clínicas negativas: la zona C podría llevar a una intervención innecesaria sobre valores de glucosa aceptables, la zona D implica una omisión de tratamiento ante una situación peligrosa (por ejemplo, hipoglucemia no detectada), y la zona E indica errores que inducirían a decisiones diametralmente opuestas a las correctas, como administrar insulina cuando no corresponde.

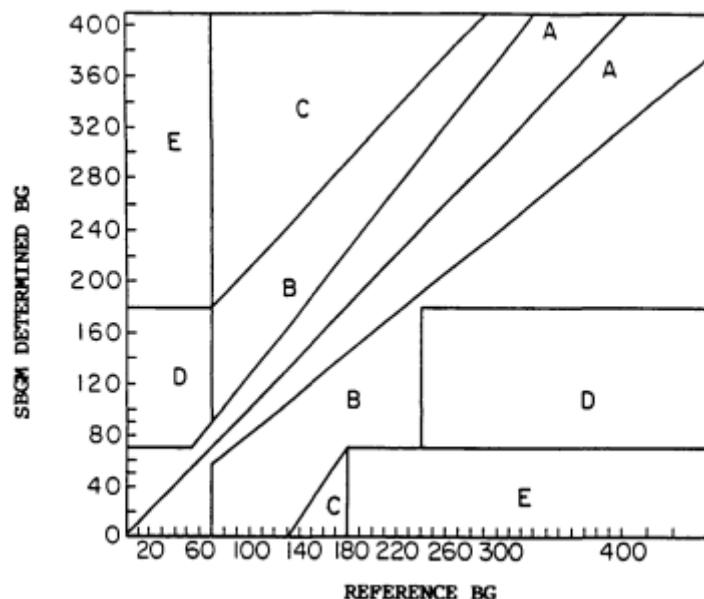


Figura 1: Extraído del paper: definición de zonas según la clasificación de Clarke

Es habitual establecer como criterio de aceptabilidad clínica que un modelo predictivo debe tener al menos el 95% de sus predicciones en las zonas A y B, y preferentemente menos del 5% en las zonas D y E, con ideal de 0% en estas últimas, lo que garantiza que el sistema no solo sea estadísticamente preciso, sino también clínicamente seguro para apoyar decisiones terapéuticas.

Sin embargo, este tipo de evaluaciones no se han llevado a cabo en este experimento por varios motivos. Por un lado, este tipo de métricas proponen márgenes de error no específicos para la predicción de glucosa postprandial con un horizonte de 2 horas. Adicionalmente, El Clarke Error Grid fue originalmente propuesto para evaluar dispositivos de medición de glucosa, no modelos de predicción a futuro, por lo que la utilidad clínica de las predicciones podría no estar correctamente respaldada por esta métrica.

En otros trabajos como “Comparative assessment of glucose prediction models for patients with type 1 diabetes mellitus applying sensors for glucose and physical activity monitoring, K. Zarkogianni, K. Mitsis, E. Litsa, et al.” [14] o “Machine learning-based glucose prediction with use of continuous glucose and physical activity monitoring data: The Maastricht Study, W. P. T. M. van Doorn, Y. D. Foreman, N. C. Schaper, et al.” [15] se sugiere que valores de MAE entre 25 y 30 mg/dL y de RMSE entre 30 y 40 mg/dL pueden ser clínicamente aceptables.

La falta de estándares para la evaluación de sistemas de predicción de glucosa y recomendación de insulina es uno de los problemas más acusados de este tipo de investigaciones. La adaptación de análisis inspirados en el de Clarke para problemas específicos de predicción como el de este trabajo es un área de investigación todavía abierta que merece ser tratada con más profundidad.

Los parámetros para el grid-search de cada modelo pueden resumirse en estas tablas:

| TabPFN | |
|--|--|
| No aplica la búsqueda de hiperparámetros | |

Tabla 2: parámetros de TabPFN para GridSearch

| Modelo | TabNet | | |
|---------------|---|--|-----------------------------------|
| Parámetros | Ancho de la capa de predicción de decisiones | Ancho del embebido de atención para cada máscara | Momentum para batch normalization |
| Valores | 8, 16, 32 | 8, 16, 32 | 0.01, 0.02, 0.15 |
| Observaciones | Se emplea early stopping con una paciencia de 20 épocas y tamaños de batch y batch virtual de 64 y 32 respectivamente | | |

Tabla 3: parámetros de TabNet para GridSearch

| Modelo | SVM | | | |
|------------|-----------------|----------------|-----------------|------------------------------------|
| Parámetros | kernel | C | gamma | epsilon |
| Valores | rbf, polinómico | 1, 2, 4, 8, 16 | 'scale', 'auto' | 0.025, 0.05, 0.1, 0.2, 0.4, 0.8, 1 |

Tabla 3: parámetros de SVM para GridSearch

| Modelo | XGBoost | | |
|------------|--------------------|---------------------|-----------------------|
| Parámetros | máxima profundidad | learning rate | número de estimadores |
| Valores | 2, 3, 6, 12 | 0.05, 0.1, 0.3, 0.9 | 25, 50, 75, 150, 300 |

Tabla 4: parámetros de XGBoost para GridSearch

| Modelo | RandomForest | | | | |
|------------|-----------------------|--------------------|------------------------------|-----------------------------|---------------------------|
| Parámetros | número de estimadores | máxima profundidad | mínimo de muestras por split | mínimo de muestras por hoja | máximo de características |
| Valores | 25, 50, 100, 200 | None, 10, 15, 20 | 2, 5, 10 | 1, 2, 4 | 'sqrt', 'log2' |

Tabla 5: parámetros de Random Forest para GridSearch

Cabe mencionar también que el entrenamiento de los modelos se ha llevado a cabo tomando los datos de los archivos de entrenamiento de cada paciente y creando un subconjunto de entrenamiento con el 85 % de los datos y un subconjunto del 15 % de los datos para validación.

La obtención de recomendaciones de dosis de insulina y tiempo de espera se obtiene, tal y como se anticipa en la introducción, mediante la evaluación de simulaciones para distintos posibles valores. En concreto, se prueba con todos los valores posibles entre 0 y 20 unidades de insulina en intervalos de 0.1 unidades y tiempos de espera entre -15 y 15 minutos, en intervalos de 3 minutos. Los resultados se pueden representar con gráficas como la siguiente:

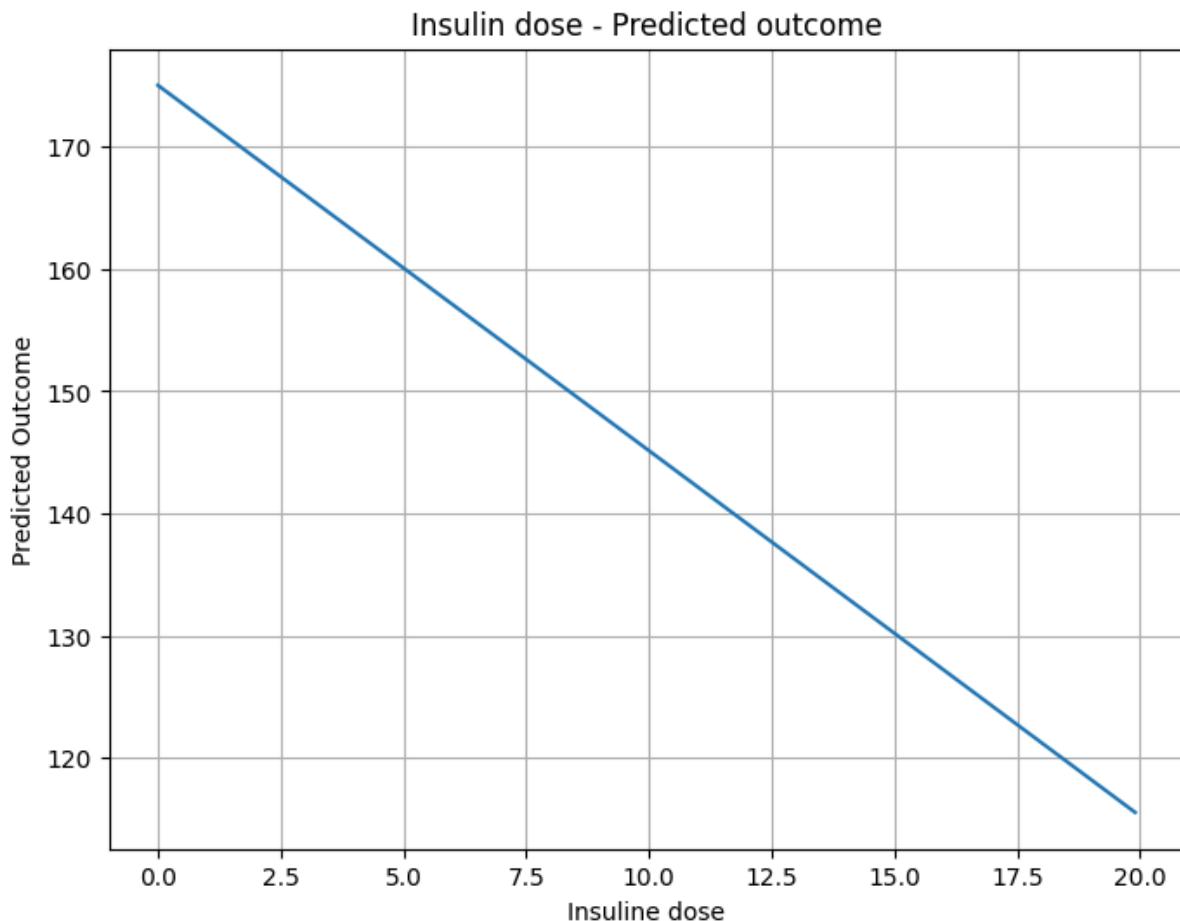


Figura 2: predicción de glucosa para distintos valores de insulina

2.4 Implementación

Toda la implementación de los algoritmos de construcción de entrenamientos, análisis de datos, preprocesado, entrenamiento de modelos y demás secciones que en este trabajo se describen, se ha llevado a cabo mediante un cuaderno en formato IPYNB de Google Colab escrito en Python y que puede encontrarse en la siguiente dirección: **TODO: añadir link al repo del trabajo**

La elección de estas tecnologías para el desarrollo del trabajo se basa en dos motivos fundamentales:

1. Python es uno de los lenguajes de programación más utilizados en el ámbito científico y tecnológico. Especialmente, en el área del aprendizaje automático, existen numerosas librerías y recursos que facilitan el desarrollo del trabajo.
2. La plataforma Google Colab ofrece un servicio gratuito limitado para el almacenamiento y ejecución remota de archivos IPYNB que permite entrenar modelos de aprendizaje automático en tarjetas gráficas en servidores de Google, lo que facilita la experimentación con modelos grandes que requieren de mayores prestaciones computacionales.

3. Marco experimental

3.1 Análisis de los datos

Para poder explotar al máximo el dataset, es preciso realizar una fase exploratoria que sirva para conocer y entender mejor qué características tienen los datos. Tras construir y etiquetar los ejemplos de entrenamiento, se han llevado a cabo las acciones de análisis que en esta sección se detallan.

3.1.1 Valores nulos

La mayoría de campos no presentan valores nulos. Los que sí presentan algún valor nulo son los siguientes:

- Horas de sueño ⇒ 17 nulos
- Calidad del sueño ⇒ 17 nulos
- Pulsaciones ⇒ 13 nulos
- GSR (Respuesta Galvánica de la piel) ⇒ 13 nulos
- Temperatura corporal ⇒ 13 nulos
- Temperatura ambiente ⇒ 13 nulos

Como puede apreciarse, la proporción de valores nulos es muy pequeña en comparación al total. Además, todos estos campos son numéricos y permiten técnicas de imputación sencillas como la media o la mediana.

3.1.2 Análisis estadístico

En primer lugar, se graficó un diagrama de dispersión de la etiqueta frente a cada una de las características para observar qué tipo de relaciones podrían existir entre cada campo y la variable objetivo.

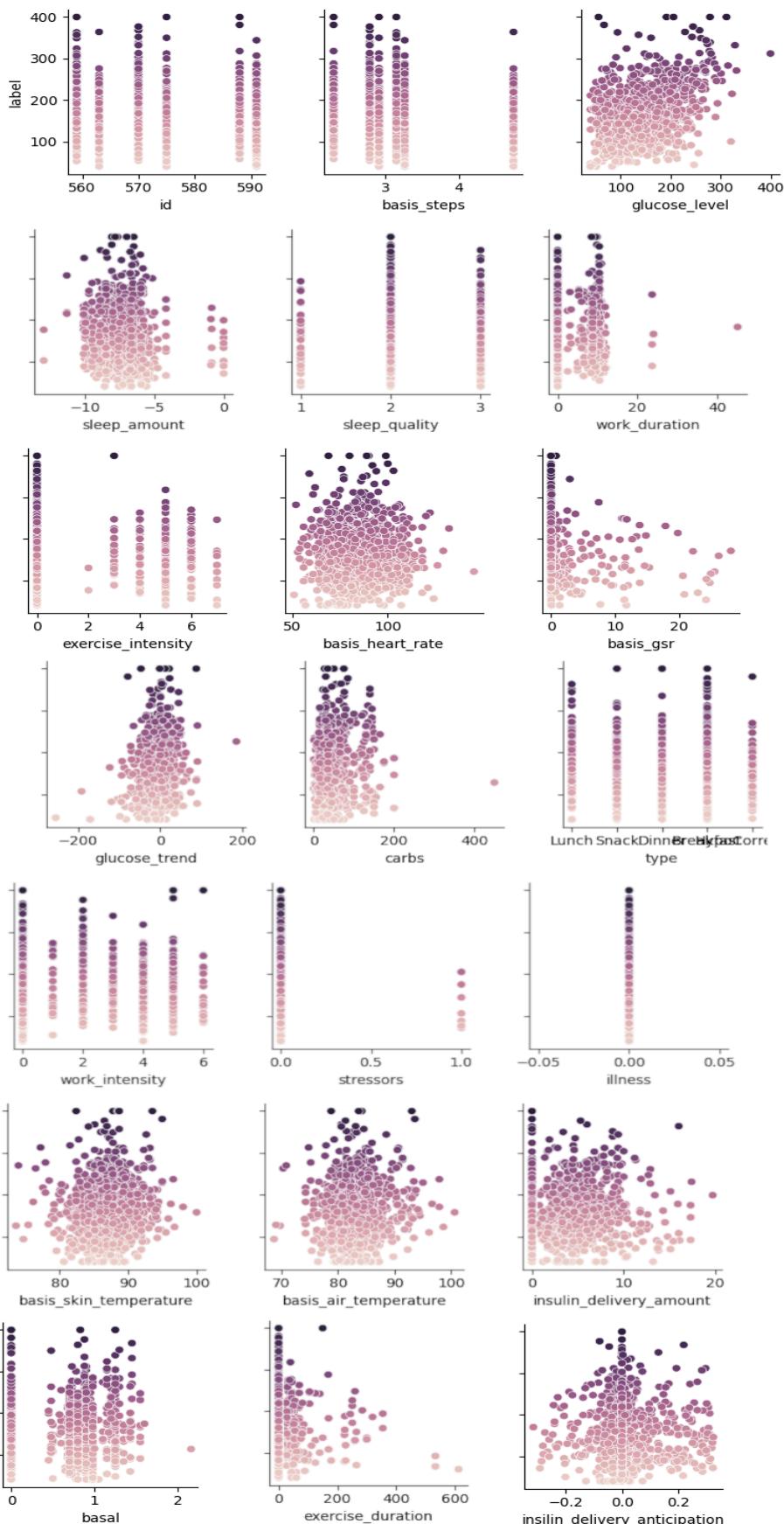


Figura 3: Diagrama de dispersión para cada característica frente a la etiqueta

Estos diagramas revelan que existe una fuerte correlación entre los niveles de glucosa previos a la ingesta y el nivel de glucosa tras ella. En concreto, valores previos más altos tienden a significar niveles posteriores más altos.

Aunque en el resto de variables no se aprecien relaciones con tanta claridad, no quiere decir que no aporten información acerca de la variable objetivo, sino que en caso de hacerlo, su relación no es lineal o no es independiente, en cuyo caso sería la variación de dos o más variables en su conjunto las que influyen en el valor de la etiqueta. Esto puede sugerir que existen relaciones complejas y no lineales, lo que requerirá de modelos que sean suficientemente expresivos para poder captarlas.

En segundo lugar, la siguiente tabla muestra diferentes estadísticas acerca de los distintos atributos:

| | Media de pasos | Glucosa pre ingesta | Tendencia de glucosa | Carbohidratos | Insulina basal | Horas de sueño |
|----------------|-----------------------|----------------------------|-----------------------------|----------------------|-----------------------|-----------------------|
| Mín. | 2.29 | 40.0 | -255.00 | 0.00 | 0.00 | 0.00 |
| Media | 3.08 | 138.6 | -6.66 | 42.5 | 0.66 | 7.31 |
| Mediana | 2.91 | 132.00 | -4.00 | 35.00 | 0.78 | 7.36 |
| Máx. | 4.74 | 400.00 | 186.00 | 450.00 | 2.16 | 13.00 |
| Desv. | 0.69 | 60.01 | 30.32 | 34.20 | 0.45 | 1.43 |

| | Calidad de sueño | Duración del trabajo | Intensidad del trabajo | Enfermedad | Duración actividad física | Intensidad actividad física |
|----------------|-------------------------|-----------------------------|-------------------------------|-------------------|----------------------------------|------------------------------------|
| Mín. | 1.00 | 0.00 | 0.00 | 0.0 | 0.00 | 0.00 |
| Media | 2.53 | 3.47 | 1.38 | 0.0 | 15.89 | 0.96 |
| Mediana | 3.00 | 0.00 | 0.00 | 0.0 | 0.00 | 0.00 |
| Máx. | 3.00 | 45.10 | 6.00 | 0.0 | 613.00 | 7.00 |
| Desv. | 0.57 | 4.91 | 1.98 | 0.0 | 51.94 | 2.00 |

| | Pulso | GSR | T. corporal | T. ambiente | Dosis de insulina | Tiempo de espera |
|----------------|--------------|------------|--------------------|--------------------|--------------------------|-------------------------|
| Mín. | 51.00 | 0.00 | 73.58 | 68.72 | 0.00 | -0.31 |
| Media | 85.87 | 0.85 | 86.85 | 83.06 | 3.32 | 0.01 |
| Mediana | 86.000 | 0.00 | 86.90 | 82.76 | 2.50 | 0.00 |
| Máx. | 145.00 | 28.00 | 100.04 | 100.58 | 19.70 | 0.32 |
| Desv. | 13.14 | 3.27 | 3.13 | 3.66 | 3.68 | 0.09 |

Tabla 6: Medidas estadísticas de las características numéricas

Hay dos detalles interesantes que se extraen de estos análisis:

1. El campo *Enfermedad* no tiene ninguna variabilidad. Una revisión de los datos en los archivos XML originales reveló que los sujetos marcaban el inicio de los períodos de enfermedad pero no el final. Por tanto, esta característica no resultará de utilidad, ya que es imposible saber qué ingestas tras el inicio de la enfermedad siguen estando bajo su efecto.
2. Algunos campos muestran valores máximos y mínimos un poco desproporcionados. Por ejemplo, la tendencia de glucosa muestra un registro de -255 puntos en 30 minutos y la cantidad de carbohidratos tiene un máximo de 450, mucho mayor que la media. Esto nos habla de la posible existencia de valores anómalos, la cual se analizará en una sección posterior.

3.2 Preprocesado de datos

El preprocesado de datos sirve para preparar el dataset para que pueda ser utilizado en el entrenamiento de modelos. Las tareas del preprocesado incluyen:

- Imputación de valores nulos.
- Codificación de variables categóricas.
- Escalado de datos.
- Eliminación de columnas no deseadas.

Mientras que los dos primeros puntos son necesarios para que los modelos de aprendizaje automático puedan procesar adecuadamente los datos, los dos siguientes puntos son técnicas que tienen como objetivo optimizar el proceso de entrenamiento.

La imputación de valores nulos se ha realizado mediante el uso de la mediana de la característica faltante. Imputar datos faltantes usando la mediana es una estrategia sólida porque la mediana es robusta frente a valores atípicos, a diferencia de otras medidas como la media, que puede distorsionarse por extremos. Esto permite mantener una representación más fiel de la distribución central de la variable.

La codificación de variables categóricas es necesaria porque los algoritmos de aprendizaje automático requieren datos numéricos para operar. La codificación transforma las variables

categóricas en valores que el modelo puede procesar, permitiendo así que se aproveche la información contenida en esas variables.

En este caso, es necesario codificar la variable *type*, que representa el tipo de comida. A la hora de codificar variables categóricas, los dos métodos más utilizados son los siguientes:

1. **Codificado ordinal:** El codificado ordinal asigna un valor entero distinto a cada categoría, preservando un posible orden implícito entre ellas (por ejemplo, bajo = 0, medio = 1, alto = 2). Es adecuado cuando la variable categórica posee una relación jerárquica o secuencial significativa, como niveles de satisfacción, clasificaciones educativas o escalas de frecuencia. Sin embargo, su uso puede inducir relaciones numéricas artificiales si el orden no es real o relevante, llevando a interpretaciones erróneas por parte del modelo.
2. **One-hot encoding:** el one-hot encoding representa cada categoría mediante un vector binario en el que solo una posición toma el valor 1 y el resto 0. Esta codificación elimina cualquier suposición de orden entre categorías, lo que la hace más apropiada para variables categóricas nominales, como colores, nombres de ciudades o tipos de comida. Si bien evita inferencias erróneas sobre relaciones ordinales, puede aumentar significativamente la dimensionalidad del conjunto de datos cuando el número de categorías es alto.

No es posible establecer un orden para esta variable debido a la existencia de los valores *snack* y *hypo correction*, que son tipos de ingestas que pueden suceder en cualquier momento del día. Por tanto, se han codificado mediante *one hot encoding*, ya que el aumento producido en la dimensionalidad no es preocupante debido a que la cantidad de valores que puede tomar esta característica no es muy elevada.

En cuanto al escalado, es fundamental porque muchos algoritmos, como los basados en distancias (por ejemplo, SVM) y los métodos de gradiente, son sensibles a la magnitud de las variables. Si los datos no se escalan, las características con valores numéricos más grandes pueden dominar el comportamiento del modelo, lo que puede conducir a un sesgo en el aprendizaje y a un rendimiento subóptimo.

La eliminación de columnas no deseadas es importante para no alimentar a los modelos con características poco informativas, como el caso de la característica *Enfermedad*. Además, como se explicará en la sección de experimentos, también se eliminarán algunas variables para tratar de obtener soluciones de una mayor practicidad de uso por los usuarios.

3.3 Detección de anomalías

La detección de anomalías (o *outliers*) permite averiguar si hay ejemplos anómalos en los datos. Que un dato sea anómalo no necesariamente significa que sea erróneo, pues a veces hay casos reales que son extremos. Sin embargo, algunos modelos pueden perder eficacia al tratar de ajustarse a este tipo de ejemplos, ya que sus características son muy distintas a las del resto y aprender los patrones que los identifican no son una garantía de generalización.

Se ha utilizado un Isolation Forest [16] para investigar la posible existencia de outliers en el dataset. Un Isolation Forest es un algoritmo de aprendizaje automático no supervisado que se utiliza para la detección de anomalías o valores atípicos. Funciona aislando las observaciones mediante la construcción aleatoria de árboles de decisión. Las observaciones que son fáciles de aislar, es decir, que requieren menos divisiones en el árbol, son consideradas anomalías. Se ha utilizado la implementación de Isolation Forest publicada en https://pyod.readthedocs.io/en/latest/_modules/pyod/models/iforest.html, que a su vez es un wrapper de la implementación de SciKit Learn. El siguiente histograma muestra la puntuación de anomalías que Isolation Forest ha asignado a los datos:

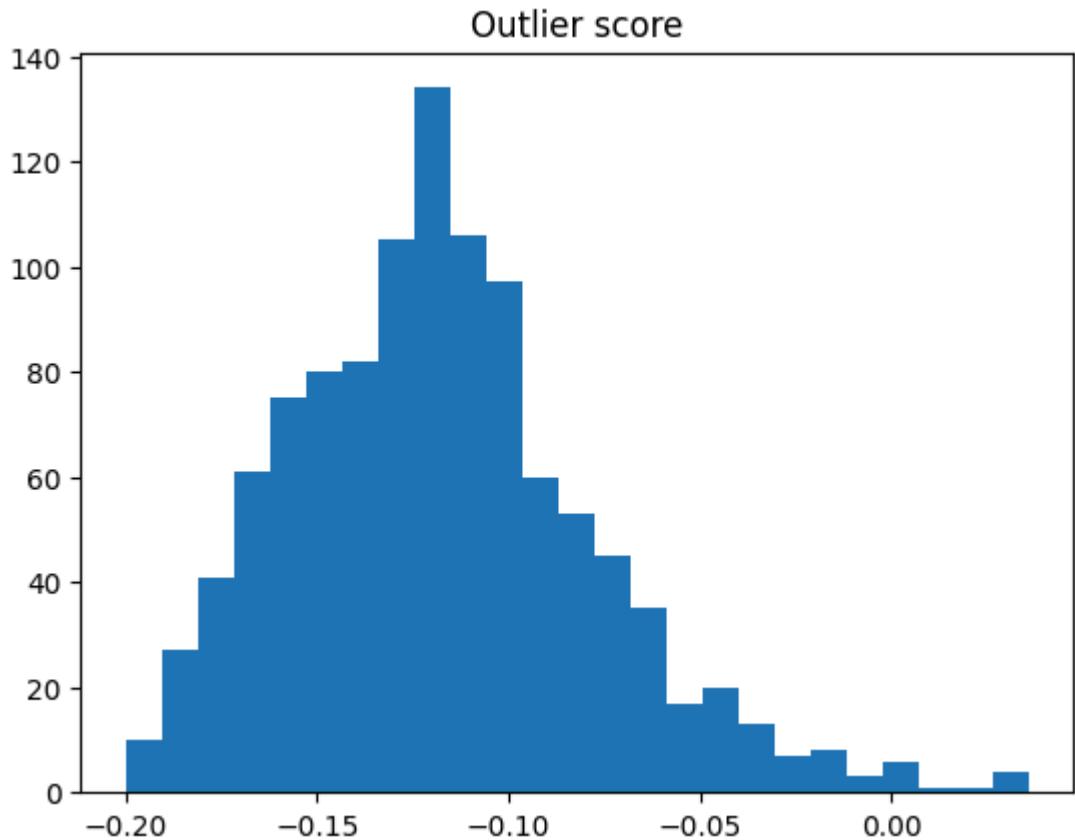


Figura 4: Histograma de puntuaciones de anomalías.

Eje x: puntuación

Eje y: cantidad de ejemplos

En el histograma se observa una forma unimodal relativamente simétrica, con un ligero sesgo hacia la derecha. La mayoría de las observaciones se concentran en un rango estrecho de valores (entre -0,15 y -0,10), lo cual sugiere una alta homogeneidad en el conjunto de datos. Aunque existe una ligera cola que se extiende hacia valores cercanos a cero e incluso positivos, esta no presenta una densidad significativa ni es evidencia de una segunda acumulación diferenciada.

En contextos donde existen verdaderos grupos de observaciones anómalas, es común observar una ruptura abrupta en la distribución, seguida de un segundo pico o cúmulo de valores anómalos, algo que no se manifiesta claramente en este caso.

La continuidad en la distribución sugiere que no existe una separación estructural evidente entre un grupo principal de datos normales y un subconjunto atípico. Es decir, o bien prácticamente no hay outliers, o bien los hay pero no son tan distintos en comparación con los ejemplos normales. Por tanto, no parece justificado tratar de aislarlos y eliminarlos, ya que no hay una señal estadística clara que respalte su existencia.

3.4 Establecimiento de una referencia

El establecimiento de una referencia consiste en obtener una serie de medidas de rendimiento a partir de otros modelos o estrategias para facilitar la interpretación de los resultados de los verdaderos modelos mediante la comparación con esta referencia.

En este caso, el objetivo es utilizar un modelo sencillo, rápido de entrenar, menos potente y no optimizado para determinar unos resultados base que sirvan para obtener una imagen del rendimiento mínimo esperado para el resto de modelos.

Es decir, si los modelos escogidos en el apartado 2.2 son teóricamente más precisos para este tipo de problema y su entrenamiento va a ser optimizado, sus resultados deberían estar por encima de la referencia por un margen considerable.

Para este trabajo, se ha escogido un modelo de regresión lineal como referencia. En concreto, se ha utilizado la implementación de la librería SciKit Learn: https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html. Este tipo de modelos asumen una relación lineal entre las variables independientes y la variable objetivo. La utilidad de esta elección reside en que una aproximación tan elemental al problema facilita la puesta en valor de la capacidad de los modelos más complejos para captar relaciones no triviales en los datos.

El entrenamiento del modelo de regresión lineal y su posterior evaluación dieron los siguientes resultados (las gráficas de las predicciones en evaluación pueden encontrarse en el anexo):

| Resultados de regresión lineal (evaluación) | |
|---|-------|
| MAE | 45.82 |
| RMSE | 57.23 |

Tabla 7: Rendimiento en evaluación del modelo de regresión lineal

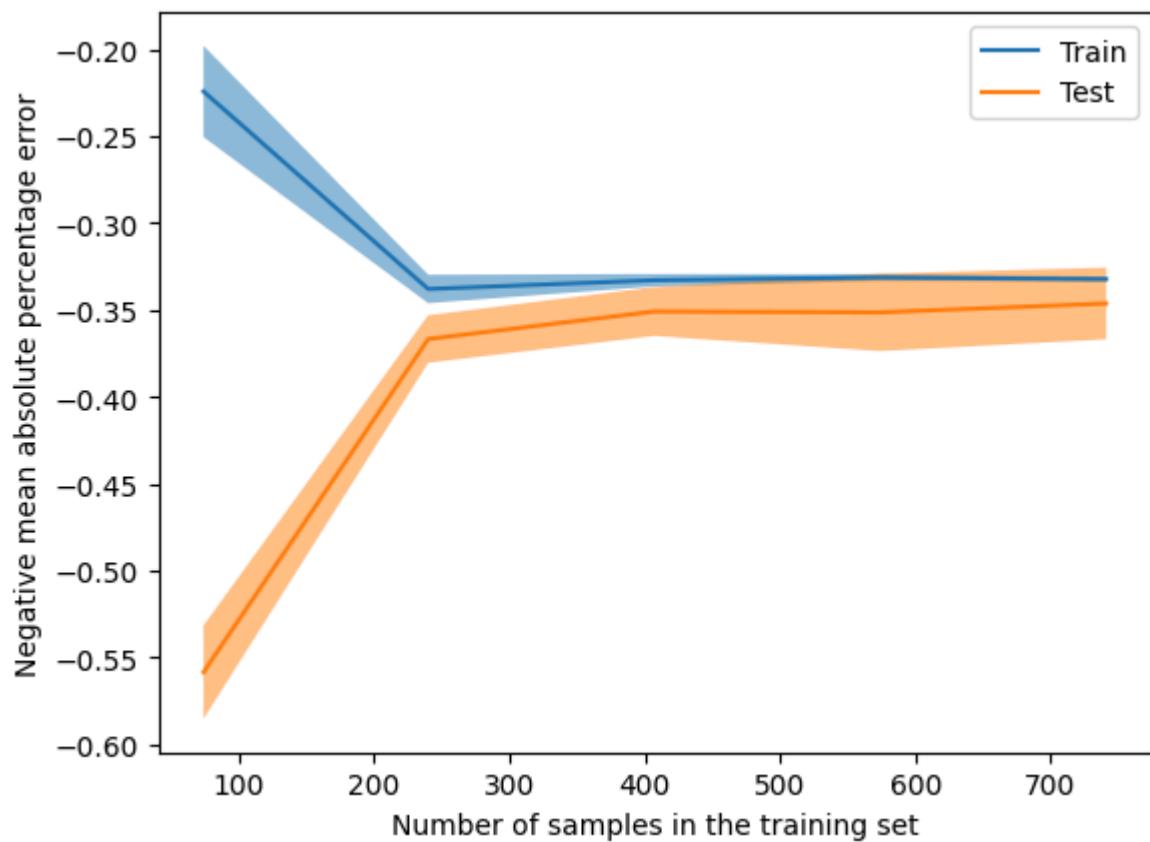
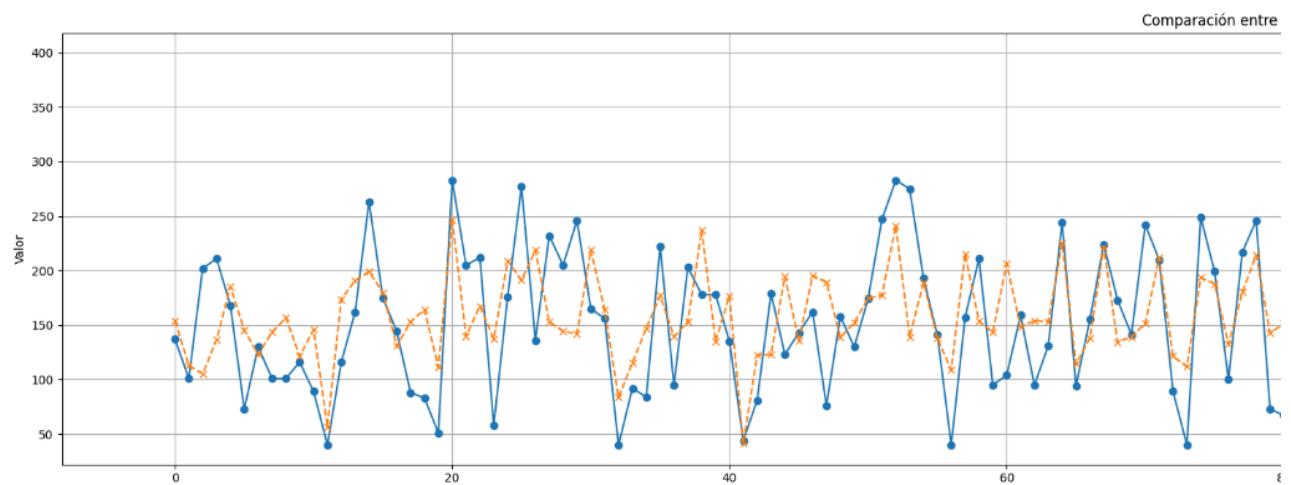


Figura 5: curvas de entrenamiento para el modelo de regresión lineal



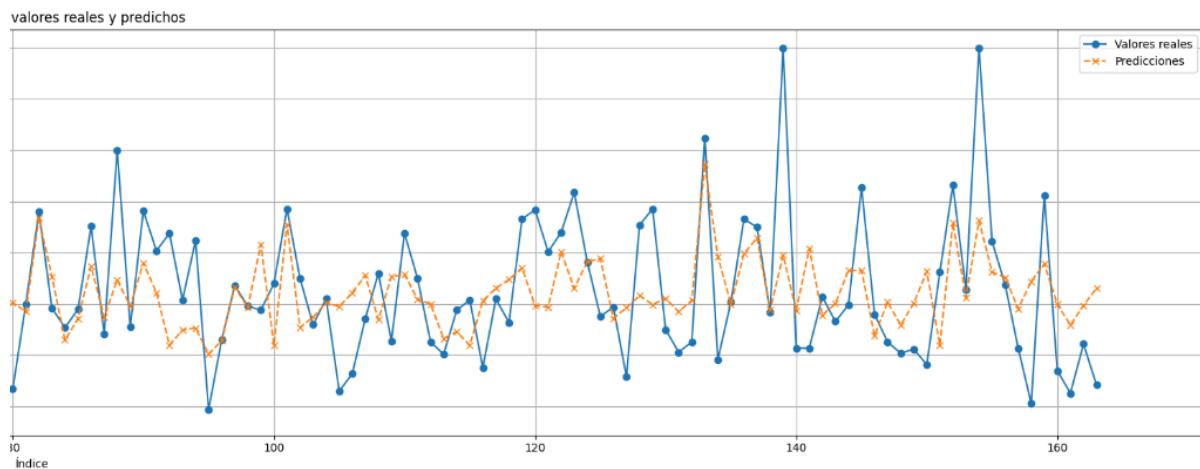


Figura 6: Predicciones del modelo de regresión lineal en evaluación

De las métricas de rendimiento y la gráfica de predicciones se extraen las siguientes conclusiones:

1. El modelo es sustancialmente peor que los modelos del estado del arte presentados en el apartado 1.3.
2. La gráfica de delata la existencia de ejemplos en los que el error cometido es muy alto. En concreto, el error es en general más pronunciado en casos donde la glucosa toma valores más bajos o más altos de lo normal, sugiriendo que el modelo podría estar prediciendo poco más que la media.

Las curvas de entrenamiento también ofrecen información reveladora. Las curvas de entrenamiento son representaciones gráficas que muestran cómo evoluciona el rendimiento de un modelo de aprendizaje automático durante el proceso de entrenamiento. Estas curvas trazan una métrica de error tanto para el conjunto de entrenamiento como para el de validación, en función del número de iteraciones, épocas o muestras vistas. Su utilidad principal radica en permitir el diagnóstico del comportamiento del modelo, facilitando la detección de fenómenos como el sobreajuste (cuando el error en entrenamiento disminuye mientras el de validación aumenta) o el subajuste (cuando ambos errores permanecen elevados).

Por un lado, las curvas de entrenamiento y validación convergen con bastante cercanía, lo que indica poco sobreajuste y un entrenamiento adecuado. Por otro lado, llama la atención que ambas curvas se aplanan tanto antes de haber visto 300 ejemplos de entrenamiento, indicando que el modelo deja de aprender mucho antes de haber visto el dataset completo. Sin embargo, puede entenderse tomando en consideración que la regresión lineal es un tipo de modelo con poca capacidad expresiva. Es esperable que los patrones de glucosa, que son tan dependientes del contexto fisiológico de las personas, presenten relaciones altamente no lineales. En este contexto, no es raro que la regresión lineal no pueda aprender mucho más a partir de los 300 ejemplos.

Las conclusiones extraídas refuerzan la idea de que la predicción de glucosa es una tarea compleja, independientemente de si se trata como un fin en sí misma o como un apoyo para

predecir dosis de insulina, como en este trabajo. La necesidad de utilizar modelos con mayor capacidad expresiva se torna aún más evidente.

4. Experimentación

4.1 Experimento 1: Modelos generales con un subconjunto de características

La mejor solución a la que puede aspirar este trabajo sería la obtención de un modelo que pueda hacer buenas recomendaciones y ser útil para cualquier usuario. Además, en un caso ideal, no haría falta tener acceso a variables poco accesibles como la GSR , las pulsaciones por minuto o la temperatura ambiente y corporal. De este modo, el sistema final no sólo sería efectivo, sino también fácil de usar.

Por tanto, los entrenamientos de este experimento cumplen con las siguientes condiciones:

- Se utiliza el dataset completo, con todos los usuarios incluidos.
- Se prescinde de las siguientes variables: nº medio de pasos, GSR, temperatura corporal y temperatura ambiente.

4.1.1 Resultados

Para las curvas de entrenamiento y las gráficas de las predicciones, véase las figuras 14 - 23 del anexo.

| Modelo | Métrica | Entrenamiento | Validación | Diferencia |
|---------------|-------------|---------------|------------|------------|
| TabPFN | RMSE | 49.45 | 52.68 | 3.23 |
| | MAE | 38.45 | 38.48 | 0.02 |
| TabNet | RMSE | 49.41 | 55.00 | 5.58 |
| | MAE | 38.54 | 45.11 | 6.57 |
| SVM | RMSE | 52.87 | 56.00 | 3.12 |
| | MAE | 39.38 | 45.57 | 6.19 |
| XGBoost | RMSE | 48.18 | 53.10 | 4.91 |
| | MAE | 37.80 | 43.23 | 5.42 |
| Random Forest | RMSE | 31.84 | 52.81 | 20.96 |
| | MAE | 24.40 | 42.88 | 18.48 |

Tabla 8: Resultados del experimento 1

Los resultados obtenidos son llamativos por varias razones. En primer lugar, los modelos han obtenido resultados mejores que la referencia propuesta en el apartado anterior, pero por una diferencia más sutil de lo que cabría esperar. En segundo lugar, todos los modelos muestran claros signos de sobreajuste, pues las métricas en evaluación empeoran significativamente respecto a las de entrenamiento. En esta línea, cabe también destacar que las curvas de entrenamiento vuelven a allanarse antes de haber visto el conjunto de entrenamiento al completo, lo que indica que el tamaño del dataset no es el cuello de botella del rendimiento de los modelos.

En términos de error, TabPFN ostenta los mejores resultados, por lo que puede considerarse el mejor modelo de este experimento. Sin embargo, a nivel de confianza clínica, ninguna de las métricas se encuentra dentro del umbral definido en la sección 2.3 (MAE entre 25 y 30 mg/dL y de RMSE entre 30 y 40 mg/dL).

4.2 Experimento 2: Modelos generales con todas las características

Debido a que el mejor modelo del experimento 1 no consigue realizar predicciones con suficiente confianza clínica, un nuevo enfoque es requerido para tratar de mejorar los resultados.

En el experimento 1, se ha prescindido de algunas de las variables disponibles para tratar de obtener un modelo más accesible para los usuarios. Sin embargo, la ausencia de estas variables ha podido ser la causa del bajo rendimiento de los algoritmos en cuanto a que hay menos información disponible para realizar predicciones.

El experimento 2 consiste en el entrenamiento de todos los modelos sin la eliminación de las variables nº medio de pasos, GSR, temperatura corporal y temperatura ambiente del dataset, con el objetivo de obtener una solución más precisa a pesar de no ofrecer tantas facilidades de uso.

4.2.1 Resultados

Para las curvas de entrenamiento y las gráficas de las predicciones, véase las figuras 24 - 33 del anexo.

| Modelo | Métrica | Entrenamiento | Validación | Diferencia |
|--------|-------------|---------------|------------|------------|
| TabPFN | RMSE | 47.07 | 51.95 | 4.88 |
| | MAE | 36.73 | 41.24 | 4.5 |
| TabNet | RMSE | 48.94 | 55.11 | 6.17 |
| | MAE | 38.77 | 43.82 | 5.05 |

| | | | | |
|----------------------|-------------|-------|-------|-------|
| SVM | RMSE | 52.35 | 56.28 | 3.92 |
| | MAE | 38.80 | 46.13 | 7.33 |
| XGBoost | RMSE | 47.55 | 52.57 | 5.02 |
| | MAE | 37.29 | 42.58 | 5.29 |
| Random Forest | RMSE | 30.90 | 52.23 | 21.33 |
| | MAE | 23.68 | 42.82 | 19.13 |

Tabla 9: Resultados del experimento 1

A pesar de contar con más características, los resultados prácticamente no han mejorado. Concretamente, XGBoost y Random Forest han mejorado muy ligeramente sus resultados, mientras que TabPFN lo ha hecho con algo más de diferencia y TabNet y SVM han sido ligeramente inferiores a sus versiones anteriores. Además, en líneas generales, los modelos sobreentrenan tanto o más que en el experimento 1.

En términos de métricas, TabPFN es de nuevo la mejor opción, pero los resultados vuelven a descartar su utilidad clínica.

4.2.2 Análisis de la importancia de las características

A pesar de contar con más información que en el experimento 1, los modelos del experimento 2 no han resultado ser sustancialmente mejores. Este hecho evidencia que las nuevas variables introducidas no son demasiado informativas del nivel de glucosa postprandial.

Entrenar modelos con características no informativas puede ser perjudicial porque añaden complejidad al problema (mayor dimensionalidad) sin ofrecer ninguna información relevante adicional para los modelos. Cabe entonces plantearse si acaso existen otras variables igualmente poco informativas.

Aprovechando que XGBoost y Random Forest ofrecen medidas de la importancia de las características de forma nativa, se ha realizado un gráfico de barras evidenciando la importancia de las características según estos modelos.

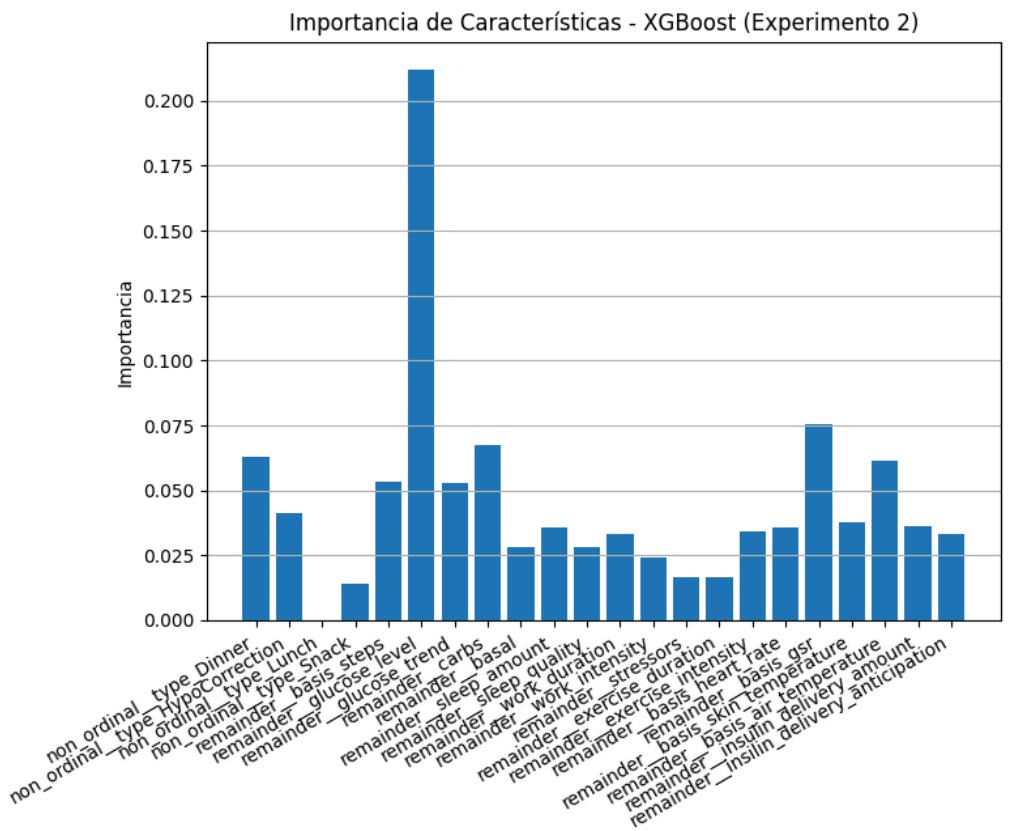


Figura 7: Importancia de las características según XGBoost

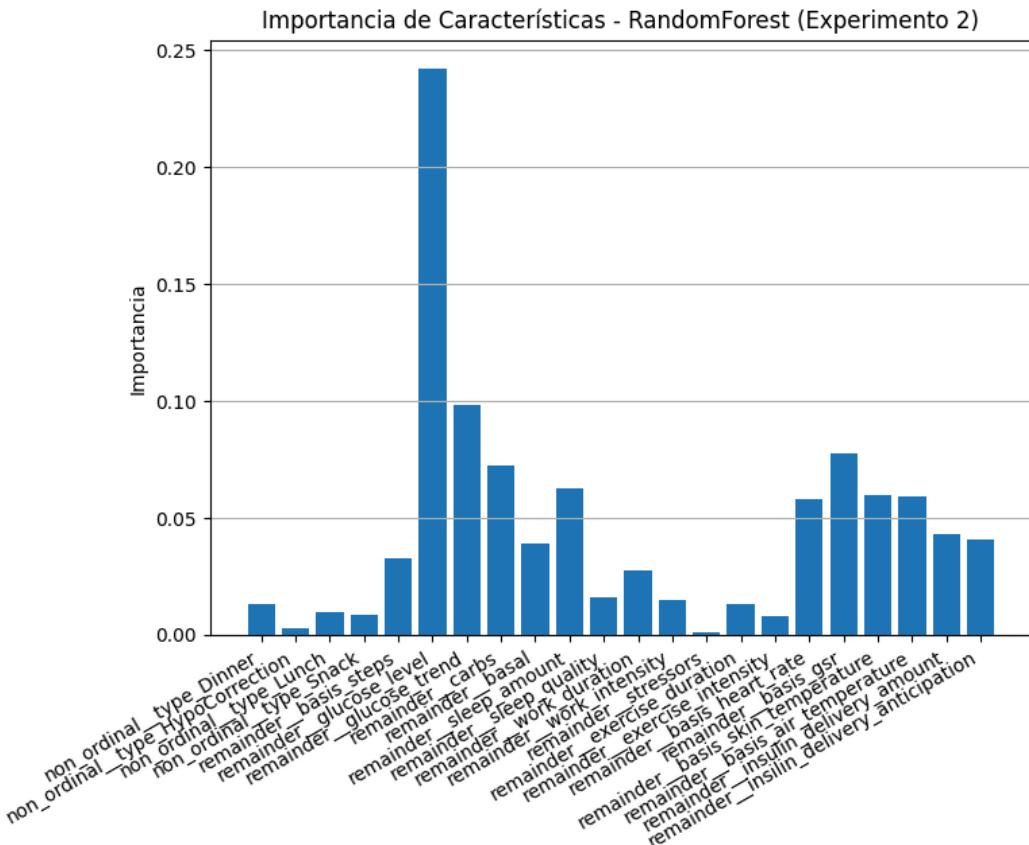


Figura 8: Importancia de las características según Random Forest

Ambos modelos coinciden en que el factor más importante, por mucha diferencia, es el nivel de glucosa previo a la ingesta. El resto de factores están mucho más equilibrados entre sí, lo que da lugar a la siguiente hipótesis: si, atendiendo a los resultados, las nuevas variables incluidas en el experimento 2 son poco informativas pero los gráficos indican que no tienen una importancia inferior al resto de variables, entonces gran parte del dataset es, en general, poco informativo.

En realidad, que la informatividad de los datos sea el cuello de botella para el rendimiento de los modelos es una hipótesis que encaja con otras observaciones realizadas en los experimentos y análisis anteriores. Al fin y al cabo, si el dataset es poco informativo:

1. Tiene sentido que los modelos no sean mucho mejores que la referencia, pues no hay más información de la que aprender a pesar de su incremento en las capacidades.
2. Es esperable que los modelos sufran de sobreentrenamiento, ya que los patrones aprendidos durante el entrenamiento no son fieles a la tendencia real de los datos y no ofrecen garantías de generalización.
3. Se explica por qué las gráficas de entrenamiento muestran que el ritmo de aprendizaje frena abruptamente antes de haber visto el dataset completo, ya que los modelos aprenden todo lo que se puede aprender antes de que el reducido tamaño del dataset sea un problema

Hay varios motivos por los que el conjunto de datos puede ofrecer pocas prestaciones. Por un lado, el número de características que los pacientes han tomado manualmente es elevado. Por ejemplo, la cantidad de carbohidratos de un alimento es crucial para la correcta predicción de los niveles de glucosa, pero la imprecisión en su registro puede estar introduciendo ruido. Por otro lado, existen variables como la calidad del sueño o intensidad del ejercicio físico que los usuarios deben introducir en base a sensaciones subjetivas, lo que puede provocar inconsistencias en su registro.

Sin embargo, hay un detalle fundamental: los modelos de los experimentos 1 y 2 utilizan los datos de todos los pacientes a la vez. La diabetes es una enfermedad en la que el metabolismo juega un papel crucial. El peso, la edad o el sexo, entre muchos otros, son factores que afectan al metabolismo y del cual no hay información en el dataset. Cabe la posibilidad de que la información de los datos sea relevante pero que los modelos no puedan inferir cómo influyen en los niveles de glucosa, por ejemplo, la cantidad de carbohidratos o la dosis de insulina, si no tienen información para aproximar cómo es el metabolismo del usuario. Además, al haber información de sólo 6 pacientes, también es posible que la muestra de diabéticos sea demasiado pequeña como para que los modelos puedan aprender cómo se comporta la diabetes en general

4.3 Experimento 3: Modelos personalizados

Con el objetivo de arrojar algo de luz acerca de la cuestión de por qué los modelos de los experimentos 1 y 2 no ofrecen predicciones con suficiente confianza clínica, este tercer experimento viene a proponer un entrenamiento personalizado para los datos de cada paciente.

Entrenar modelos de forma específica para cada usuario tiene las siguientes consecuencias:

1. Los modelos pueden tener mayor capacidad de aprendizaje, puesto que el efecto de las distintas variables sobre los niveles de glucosa debería ser más predecible si sólo se considera un paciente, tal y como se razona al final del apartado anterior.
2. Los modelos resultantes sólo garantizarían generalización para el paciente en el que ha sido entrenado, lo que impide la tipología de solución del experimento 1 en el que un único modelo general sería usable para cualquier usuario. Dado un nuevo usuario, sería necesaria la obtención de un registro de su información y el entrenamiento de una nueva instancia del modelo.

En cualquier caso, este experimento ofrece la posibilidad de disipar las dudas acerca de los datos. Si son de suficiente calidad y el problema radica en la incapacidad de los modelos para inferir el metabolismo de los usuarios en el caso general, se espera un salto notable en la capacidad predictiva de los modelos. Si el problema se halla en la calidad de los datos, al margen del metabolismo o muestra de usuarios, se esperan resultados similares a los obtenidos en los experimentos 1 y 2.

4.3.1 Resultados

Para las curvas de entrenamiento, las gráficas de las predicciones y diagramas comparativos, véase las figuras 34 - 49 del anexo.

| | | ID de los usuarios | | | | | |
|-----------------------|----------------------|--------------------|-------|-------|-------|-------|-------|
| | Modelos | 559 | 563 | 570 | 575 | 588 | 591 |
| RMSE en entrenamiento | TabPFN | 69.25 | 34.38 | 46.02 | 47.32 | 40.32 | 48.27 |
| | TabNet | 68.34 | 58.94 | 42.32 | 51.15 | 43.91 | 59.46 |
| | SVM | 69.43 | 39.49 | 56.23 | 50.16 | 43.63 | 46.84 |
| | XGBoost | 70.55 | 40.80 | 44.18 | 43.17 | 43.45 | 43.53 |
| | Random Forest | 50.79 | 24.62 | 22.63 | 30.76 | 21.12 | 28.97 |
| RMSE en validación | TabPFN | 51.34 | 66.58 | 47.14 | 59.81 | 48.38 | 53.38 |

| | | | | | | | |
|--|----------------------|-------|-------|-------|-------|-------|-------|
| | TabNet | 67.00 | 60.87 | 60.65 | 46.93 | 49.79 | 65.99 |
| | SVM | 54.01 | 58.95 | 55.91 | 64.30 | 54.91 | 60.97 |
| | XGBoost | 52.72 | 59.11 | 54.71 | 60.12 | 57.20 | 52.56 |
| | Random Forest | 53.37 | 57.49 | 52.03 | 59.63 | 52.68 | 56.21 |

Tabla 10: resultados desglosados de los modelos por paciente

| | | ID de los usuarios | | | | | |
|--------------------------------|----------------------|--------------------|------------|------------|------------|------------|------------|
| | | 559 | 563 | 570 | 575 | 588 | 591 |
| RMSE medio por paciente | Entrenamiento | 65.67 | 39.64 | 42.27 | 44.51 | 38.48 | 45.41 |
| | Validación | 55.68 | 60.6 | 54.08 | 58.15 | 52.59 | 57.82 |

Tabla 10: resultados medios para cada paciente

| | | Modelos | | | | |
|------------------------------|----------------------|---------------|---------------|------------|----------------|----------------------|
| | | TabPFN | TabNet | SVM | XGBoost | Random Forest |
| RMSE medio por Modelo | Entrenamiento | 47.59 | 54.02 | 50.96 | 47.61 | 29.815 |
| | Validación | 54.43 | 58.53 | 58.17 | 56.07 | 55.23 |

Tabla 11: resultados medios para cada modelo

Los resultados indican que la estrategia del entrenamiento individualizado no ha conducido a una mejora general del rendimiento de los modelos. En términos del promedio, el mejor modelo ha vuelto a ser TabPFN. Con un RMSE medio de 54,43 , ya no sólo se encuentra fuera del rango aceptable a nivel clínico, sino que representa un retroceso en el rendimiento respecto a su versión del experimento 2 (RMSE de 51,95).

Aunque las motivaciones para realizar entrenamientos individuales siguen siendo razonables, existen algunos motivos por los que los resultados han podido no mejorar. Por un lado, aunque se podría asumir que el rendimiento deficiente se debe a una escasez de datos por paciente, las curvas de entrenamiento indican que los modelos dejan de mejorar antes de haber recorrido el conjunto completo de entrenamiento, lo que sugiere que el cuello de botella no está en la cantidad absoluta de datos, sino en su calidad e informatividad, tal y como se anticipaba en el inicio de esta sección. Es decir, los ejemplos

disponibles por paciente, aunque suficientes en número, podrían carecer de diversidad o no contener suficiente señal predictiva. En este contexto, los entrenamientos globales pueden resultar en mejores modelos debido a que se benefician de patrones generales observados entre muchos individuos (como respuestas promedio a ingestas o tendencias fisiológicas comunes), lo que actúa como una forma de regularización implícita.

Además, se observa una amplia diferencia en los resultados obtenidos para cada usuario particular. Estas diferencias podrían explicarse por distintos factores:

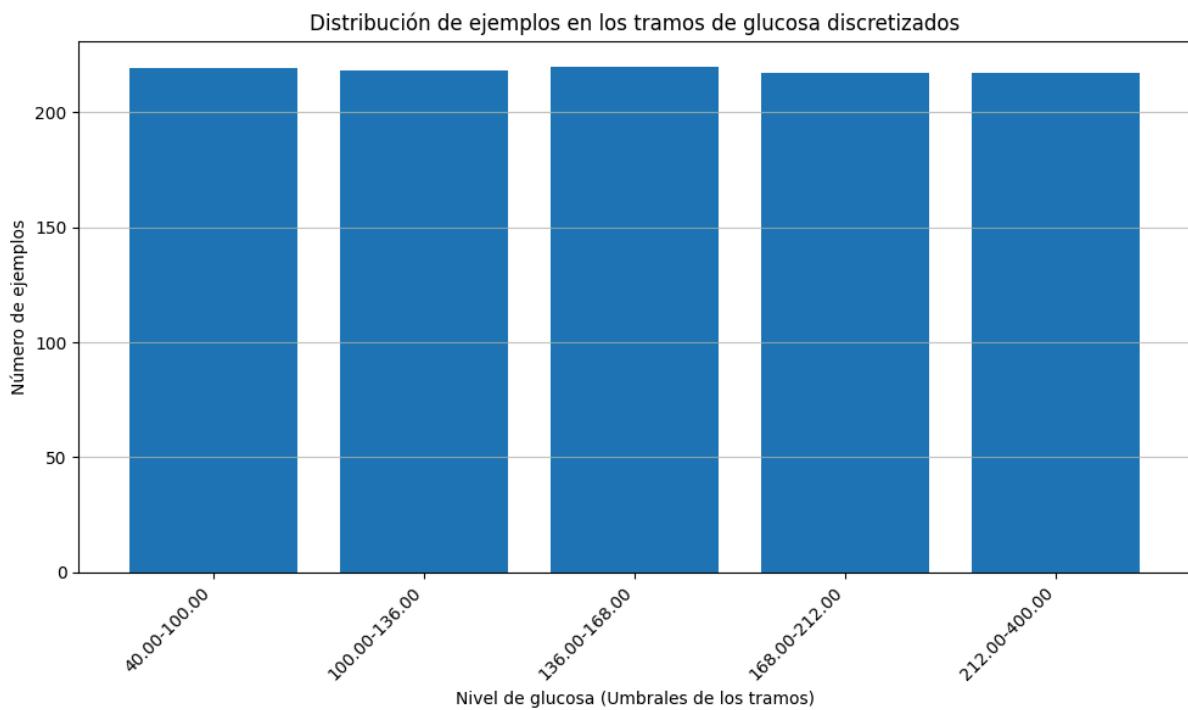
1. Heterogeneidad en el comportamiento fisiológico: Algunos pacientes pueden tener respuestas más predecibles y estables, lo que facilita el aprendizaje.
2. Calidad de los registros: Variaciones en la precisión de las anotaciones de carbohidratos, insulina y eventos relevantes pueden afectar a la calidad del dataset individual.
3. Diversidad real de los datos: Aunque el número de ejemplos es similar, no todos los pacientes tienen la misma riqueza de contexto y variabilidad en sus datos.

En conjunto, estos hallazgos comprometen la posibilidad de construir un sistema capaz de recomendar dosis de insulina con la suficiente confianza clínica a partir de modelos predictores de glucosa entrenados con este conjunto de datos.

4.4 Experimento 4: Modelos generales con discretización

En un último intento por conseguir un sistema que pueda ofrecer asesoramiento con suficiente confianza clínica, este experimento viene a proponer un nuevo enfoque en el que se sacrifica precisión en las recomendaciones a cambio de ofrecer mayor seguridad en las mismas.

En concreto, se ha llevado a cabo una reformulación del problema como una tarea de clasificación multiclase. La nueva formulación divide los valores de glucosa en cinco clases discretas, definidas a partir de percentiles, quedando los datos etiquetados según la siguiente distribución:



- Clase 0: 40 - 100 mg/dL
- Clase 1: 100 - 136 mg/dL
- Clase 2: 136 - 168 mg/dL
- Clase 3: 168 - 212 mg/dL
- Clase 4: 212 - 400 mg/dL

La elección de los rangos con base en una distribución aproximadamente uniforme permite entrenar modelos más estables, evitando un potencial desbalanceo de las clases, lo que facilita la generalización y la sensibilidad de los modelos.

La elección de 5 clases con estos rangos, pese a parecer artificial, puede justificarse a nivel clínico. Las clases 1 y 2 están compuestas únicamente por casos normoglucémicos, mientras que el resto de clases contienen en su mayoría o exclusivamente casos de hipoglucemia (clase 0) o hiperglucemia (clases 3 y 4). De este modo, es posible construir un sistema de asesoramiento en el que se recomienda al usuario aquel rango de dosis de insulina que mantenga al usuario bien en el rango 1 o en el rango 2 de glucosa pasadas 2 horas tras la ingesta.

Reformular el problema de predicción de glucosa como una tarea de clasificación en lugar de regresión puede resultar en modelos más precisos debido a que reduce la complejidad del espacio de salida y limita el impacto de errores pequeños que, en regresión, pueden ser penalizados desproporcionadamente. Al discretizar los valores de glucosa en clases bien definidas, el modelo no necesita estimar un valor exacto —lo cual es especialmente difícil en presencia de ruido fisiológico y mediciones imprecisas—, sino que solo debe identificar el rango correcto, una tarea más robusta frente a la variabilidad interindividual e intrapersonal.

Por otro lado, surgen de manera natural formas de evaluar y entender cómo predicen los modelos. Utilizando matrices de confusión, es posible ver de qué manera falla el modelo y

razonar si esos fallos son o no tolerables. Por ejemplo, no es lo mismo fallar entre las clases 1 y 2 que entre las clases 1 y 3. De este modo, se puede escoger qué modelo es mejor con mayor criterio y no sólo en base al porcentaje de aciertos.

4.4.1 Resultados

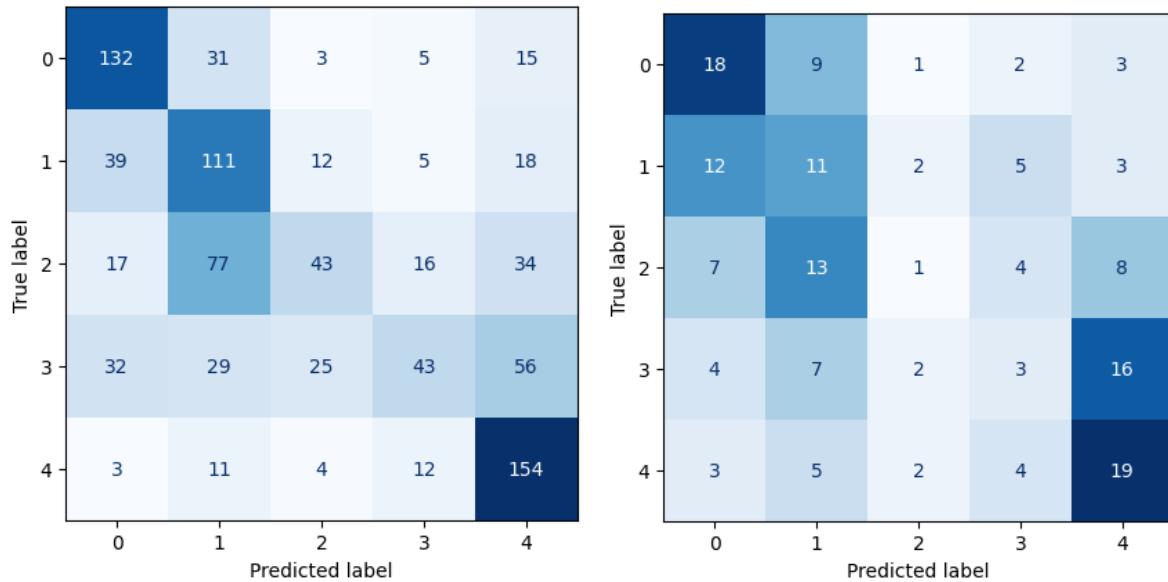


Figura 9: Matrices de confusión para TabPFN en entrenamiento (izquierda) y evaluación (derecha)

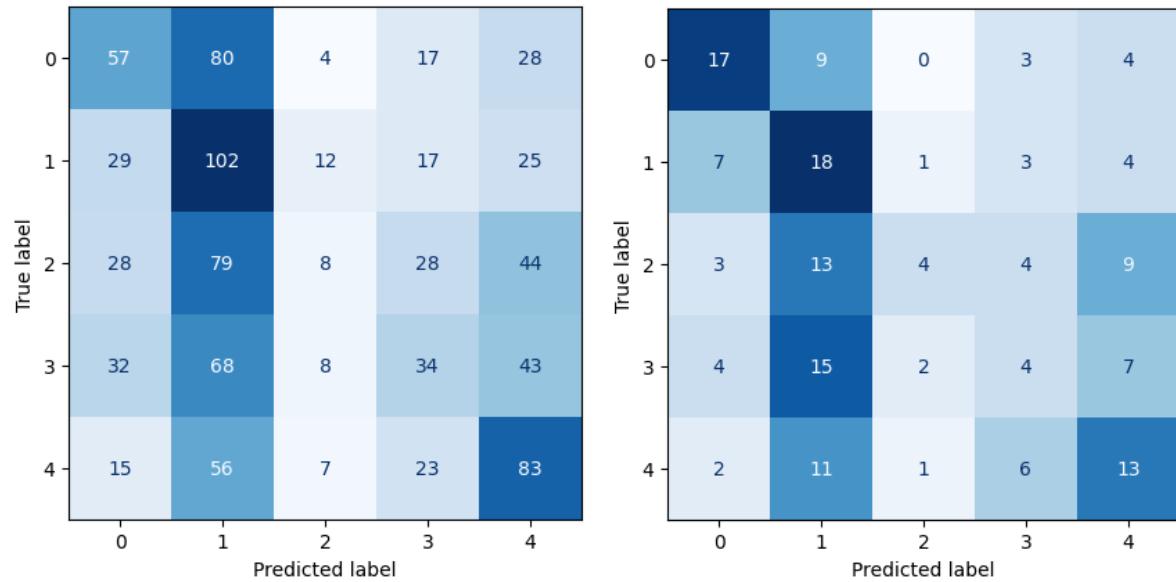


Figura 10: Matrices de confusión para TabNet en entrenamiento (izquierda) y evaluación (derecha)

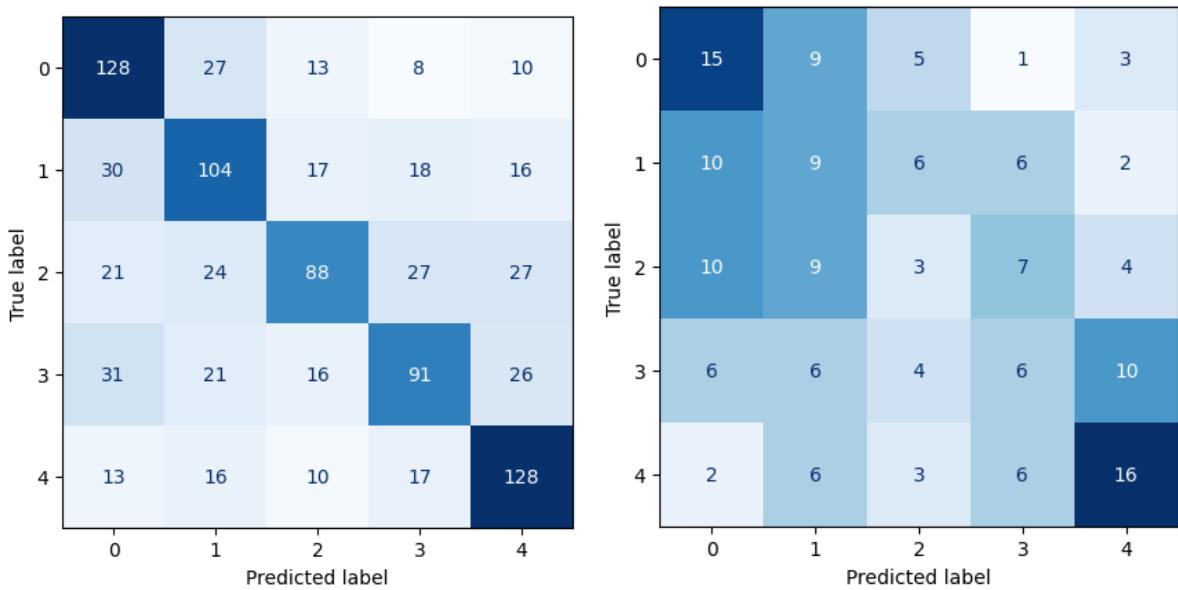


Figura 11: Matrices de confusión para SVM en entrenamiento (izquierda) y evaluación (derecha)

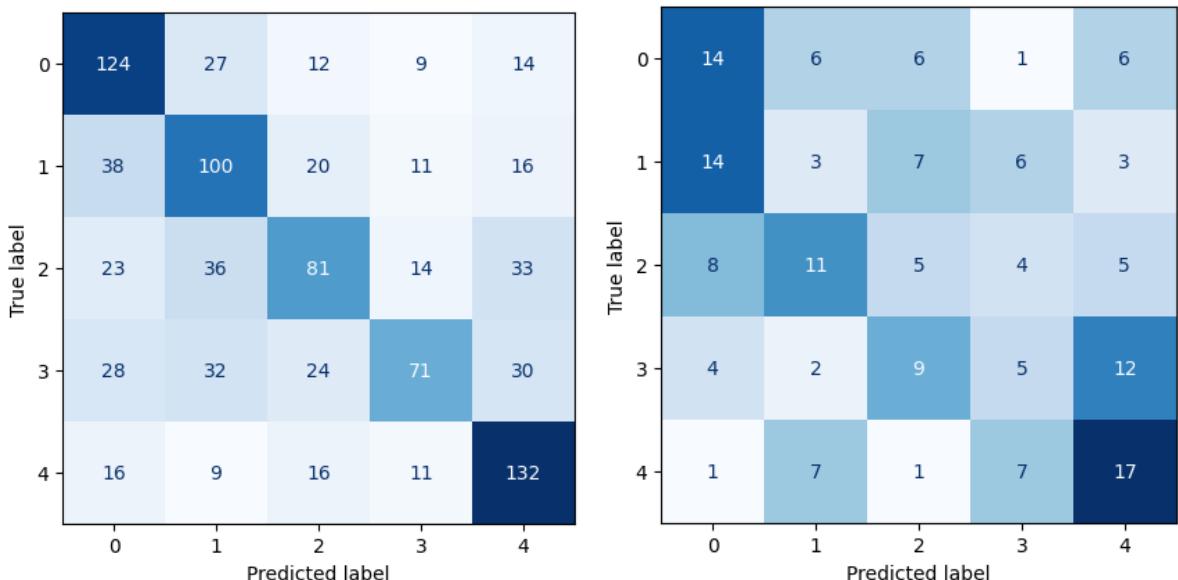


Figura 12: Matrices de confusión para XGBoost en entrenamiento (izquierda) y evaluación (derecha)

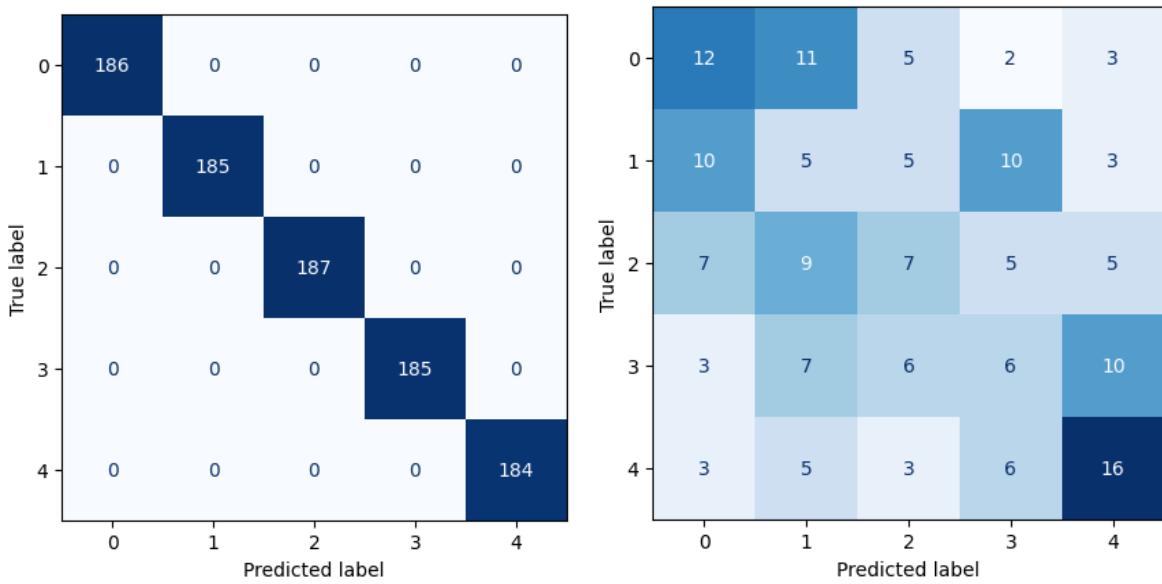


Figura 13: Matrices de confusión para Random Forest en entrenamiento (izquierda) y evaluación (derecha)

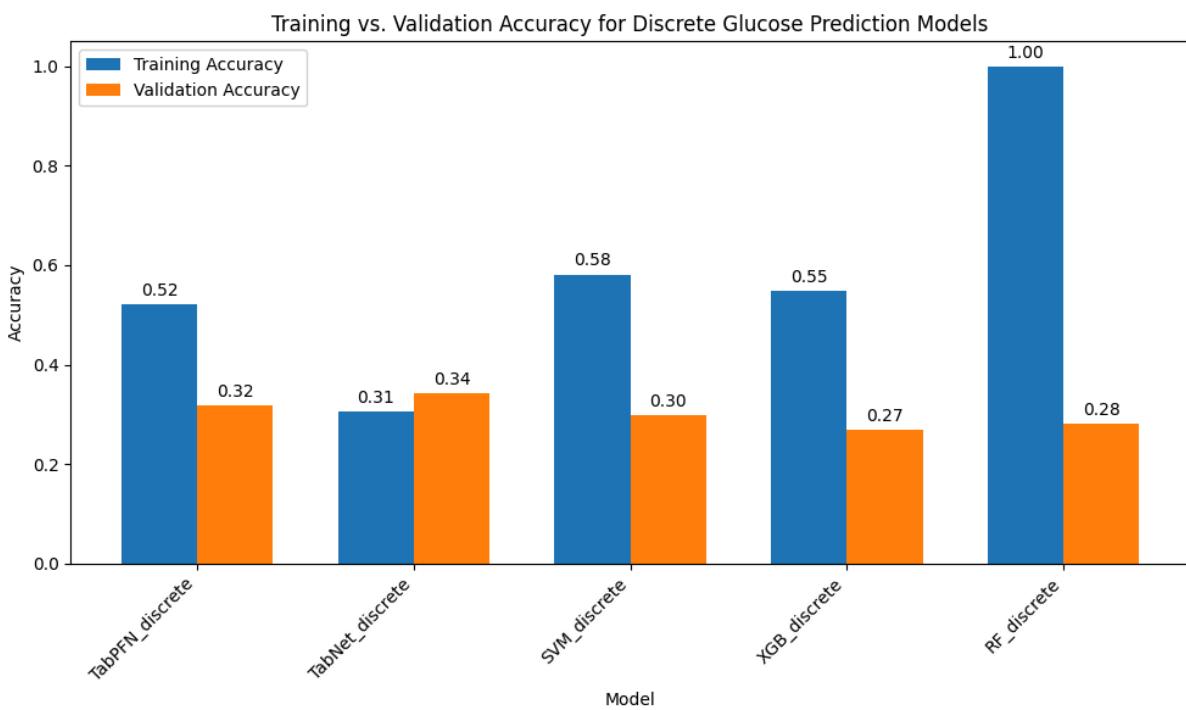


Figura 14 :Accuracy por modelo

Los resultados de este experimento indican que este nuevo enfoque es insuficiente para dotar a los modelos de fiabilidad en sus predicciones. No solo se obtienen predicciones muy por debajo del 50% para todos los modelos en evaluación, sino que se observa una gran dificultad para etiquetar precisamente las clases 1 y 2, que son las más importantes para este enfoque.

5. Análisis de los resultados

A lo largo de este trabajo, diversos modelos han sido puestos a prueba para evaluar su capacidad para predecir los niveles de glucosa de diabéticos de tipo 1 pasadas 2 horas tras la ingesta de alimentos.

En el primer experimento, los modelos fueron entrenados utilizando un conjunto reducido de características y usando los datos de todos los pacientes con el objetivo de obtener una solución versátil y práctica. El bajo rendimiento de dichos modelos propició el reentrenamiento de los mismos utilizando esta vez todas las características disponibles de los datos. Pese al aumento en el número de características, el rendimiento no mejoró, lo que dió lugar a un entrenamiento individualizado por paciente en el experimento 3. Una vez más, los resultados situaron a los modelos fuera del umbral de confianza clínica, lo que derivó en un cambio de enfoque para el experimento 4, en el que se llevó a cabo una transformación en el etiquetado de los datos para pasar de un problema de regresión a uno de clasificación. Los resultados fueron similares a los de los experimentos anteriores.

Cada uno de los modelos utilizados realizan predicciones de formas distintas y presentan arquitecturas únicas. Entre los modelos seleccionados, se encuentran algunos de los más utilizados en el ámbito de los problemas tabulares, como XGBoost, y propuestas más innovadoras y rompedoras como TabPFN, que ya ha demostrado mejorar el rendimiento de muchos de los modelos más punteros del estado del arte en una amplia gama de problemas similares [17]. Además, cada experimento propone una situación de aprendizaje adaptada al conocimiento extraído del experimento anterior.

Aún con todas estas consideraciones, los resultados no alcanzan en ningún caso los umbrales requeridos para ser considerados aptos para situaciones clínicas reales. Por tanto, si hay variedad de modelos, experimentos y aproximaciones pero los resultados no varían demasiado, la hipótesis más razonable es que no es posible obtener predicciones significativamente mejores que las obtenidas en este trabajo empleando únicamente esta construcción de datos a partir del dataset Ohio T1DM 2018.

En cualquier caso, los resultados no son excesivamente inferiores a otros obtenidos con el dataset Ohio T1DM (como los presentados en la sección 1.3) tomando en consideración que los modelos aquí mostrados predicen específicamente niveles de glucosa postprandiales, tarea especialmente difícil debido a la compleja influencia del metabolismo humano en contextos de ingestas, demostrando así la utilidad del enfoque en la construcción de situaciones de ingesta que este trabajo propone.

6. Conclusiones

El objetivo inicial de este trabajo ha sido la construcción de un sistema de asesoramiento para asistir a diabéticos de tipo 1 a decidir qué dosis de insulina y tiempo de espera deben hacer ante una ingesta. No obstante, la falta de confianza clínica en los modelos predictores de glucosa ha obligado a descartar parte de esta idea, puesto que no tiene sentido crear un sistema de recomendación si los modelos que lo sustentan no son fiables. De este modo, el esfuerzo de la investigación se ha ido trasladando hacia la optimización de modelos hasta que el problema de la falta de información relevante en el conjunto de datos utilizado ha evidenciado un tope teórico para las herramientas disponibles para este trabajo. A pesar de no haber concluido con los resultados esperados, esta investigación está lejos de haber supuesto un fracaso.

Por un lado, determinar cuáles son las limitaciones de los modelos y los datos con los que se trabaja resulta crucial para tomar decisiones informadas acerca de los siguientes pasos que investigaciones similares pueden tomar. Concretamente, este trabajo pone de manifiesto que los siguientes esfuerzos deben ir en el camino de aumentar la informatividad de los datos utilizados, ya sea incluyendo registros de otras fuentes o construyendo ejemplos de forma distinta.

En esta línea, también destaca la importancia de dotar a los datos de características que definan con más precisión el metabolismo de los pacientes. Por ejemplo, adicionalmente a la cantidad de horas que una persona ha dormido en un día, se podría contemplar la cantidad de horas media que el usuario ha dormido por día entre todos los registros. Técnicas como estas podrían ayudar a dibujar con más precisión cuál es el comportamiento de la glucosa frente a distintos perfiles de usuarios. Adicionalmente, queda abierta la puerta a la exploración de técnicas de *transfer learning* en las que se toma un modelo preentrenado en un conjunto grande de usuarios y luego se aplica sobre él un entrenamiento más superficial con datos de un sólo usuario en el que deberá especializarse, combinando potencialmente los beneficios de los enfoques generalistas y especialistas.

Por otro lado, este trabajo demuestra la necesidad de investigar más acerca de técnicas de predicción de glucosa y asesoramiento de insulina y tiempos de espera. Según datos de la IDF (*International Diabetes Federation*), más de 9 millones de personas padecen esta enfermedad que no se puede prevenir ni curar [18]. La calidad de vida de estas personas depende de los avances científicos en la materia, muchos de los cuales requieren de información y registros de calidad acerca de la interacción con su enfermedad. Iniciativas como la libre publicación del dataset Ohio T1DM son cruciales en este sentido y urge que otras organizaciones realicen esfuerzos similares (respetando siempre los acuerdos de privacidad de los datos de las personas).

A medida que más investigaciones publican sus resultados en la materia, más crítico es también el establecimiento de medidas de rendimiento que aporten valor desde el punto de vista médico. Tal y como se ha comentado en el apartado 2.3, técnicas como el análisis de

error de Clarke son fundamentales y su extrapolación a otros casos de uso de la predicción de glucosa en diabéticos de tipo 1 es menester.

7. Referencias

- [1] B. G. Katzung, *Basic & Clinical Pharmacology*, 9th ed., New York, NY, USA: McGraw-Hill, 2007, ch. 41.
- [2] J. Massana, F. Torrent-Fontbona, and B. López, “Insulin recommender systems for T1DM: A review,” *Advances in Experimental Medicine and Biology*, vol. 1307, pp. 331–355, 2021. doi: [10.1007/5584_2020_482](https://doi.org/10.1007/5584_2020_482).
- [3] M. A. Karagoz, M. D. Breton, and A. El Fathi, “A comparative study of transformer-based models for multi-horizon blood glucose prediction,” *arXiv preprint arXiv:2505.08821*, 2025. [Online]. Available: <https://arxiv.org/abs/2505.08821>
- [4] M. De Bois, M. A. E. Yacoubi, and M. Ammi, “GLYFE: review and benchmark of personalized glucose predictive models in type 1 diabetes,” *Medical & Biological Engineering & Computing*, vol. 60, no. 1, pp. 1–17, 2022. doi: [10.1007/s11517-021-02437-4](https://doi.org/10.1007/s11517-021-02437-4)
- [5] C. Marling and R. Bunescu, “The OhioT1DM dataset for blood glucose level prediction,” in *Proc. 3rd Int. Workshop on Knowledge Discovery in Healthcare Data*, Stockholm, Sweden, Jul. 2018. [Online]. Available: <http://ceur-ws.org/Vol-2148/paper09.pdf>
- [6] Google, “Google Colaboratory,” [Online]. Available: <https://colab.research.google.com/>. [Accessed: Jun. 13, 2025].
- [7] N. Hollmann, S. Müller, L. Purucker, *et al.*, “Accurate predictions on small data with a tabular foundation model,” *Nature*, vol. 637, pp. 319–326, 2025. doi: [10.1038/s41586-024-08328-6](https://doi.org/10.1038/s41586-024-08328-6)
- [8] S. O. Arik and T. Pfister, “TabNet: Attentive interpretable tabular learning,” *arXiv preprint arXiv:1908.07442*, 2020. [Online]. Available: <https://arxiv.org/abs/1908.07442>
- [9] C. Cortes and V. Vapnik, “Support-vector networks,” *Machine Learning*, vol. 20, pp. 273–297, 1995. doi: [10.1007/BF00994018](https://doi.org/10.1007/BF00994018)
- [10] F. Pedregosa, G. Varoquaux, A. Gramfort, *et al.*, “Scikit-learn: Machine learning in Python,” *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011. [Online]. Available: <https://scikit-learn.org/stable/> [Accessed: Jun. 13, 2025].
- [11] T. Chen and C. Guestrin, “XGBoost: A scalable tree boosting system,” in *Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining*, San Francisco, CA, USA, Aug. 2016, pp. 785–794. doi: [10.1145/2939672.2939785](https://doi.org/10.1145/2939672.2939785).
- [12] L. Breiman, “Random forests,” *Machine Learning*, vol. 45, pp. 5–32, 2001. doi: [10.1023/A:1010933404324](https://doi.org/10.1023/A:1010933404324)

- [13] W. L. Clarke, D. Cox, L. A. Gonder-Frederick, W. Carter, and S. L. Pohl, "Evaluating clinical accuracy of systems for self-monitoring of blood glucose," *Diabetes Care*, vol. 10, no. 5, pp. 622–628, Sep. 1987. doi: [10.2337/diacare.10.5.622](https://doi.org/10.2337/diacare.10.5.622)
- [14] K. Zarkogianni, K. Mitsis, E. Litsa, *et al.*, "Comparative assessment of glucose prediction models for patients with type 1 diabetes mellitus applying sensors for glucose and physical activity monitoring," *Medical & Biological Engineering & Computing*, vol. 53, pp. 1333–1343, 2015. doi: [10.1007/s11517-015-1320-9](https://doi.org/10.1007/s11517-015-1320-9)
- [15] W. P. T. M. van Doorn, Y. D. Foreman, N. C. Schaper, *et al.*, "Machine learning-based glucose prediction with use of continuous glucose and physical activity monitoring data: The Maastricht Study," *PLOS ONE*, vol. 16, no. 6, p. e0253125, 2021. doi: [10.1371/journal.pone.0253125](https://doi.org/10.1371/journal.pone.0253125)
- [16] F. T. Liu, K. M. Ting, and Z.-H. Zhou, "Isolation Forest," in *Proc. 2008 Eighth IEEE Int. Conf. Data Mining*, Pisa, Italy, 2008, pp. 413–422. doi: 10.1109/ICDM.2008.17.
- [17] Q. Zhang, Y. S. Tan, Q. Tian, and P. Li, "TabPFN: One model to rule them all?," *arXiv preprint arXiv:2505.20003*, 2025. [Online]. Available: <https://arxiv.org/abs/2505.20003>
- [18] International Diabetes Federation, "Diabetes tipo 1," [Online]. Available: <https://idf.org/es/about-diabetes/types-of-diabetes/type-1-diabetes/>. [Accessed: Jun. 13, 2025]

8. Anexo

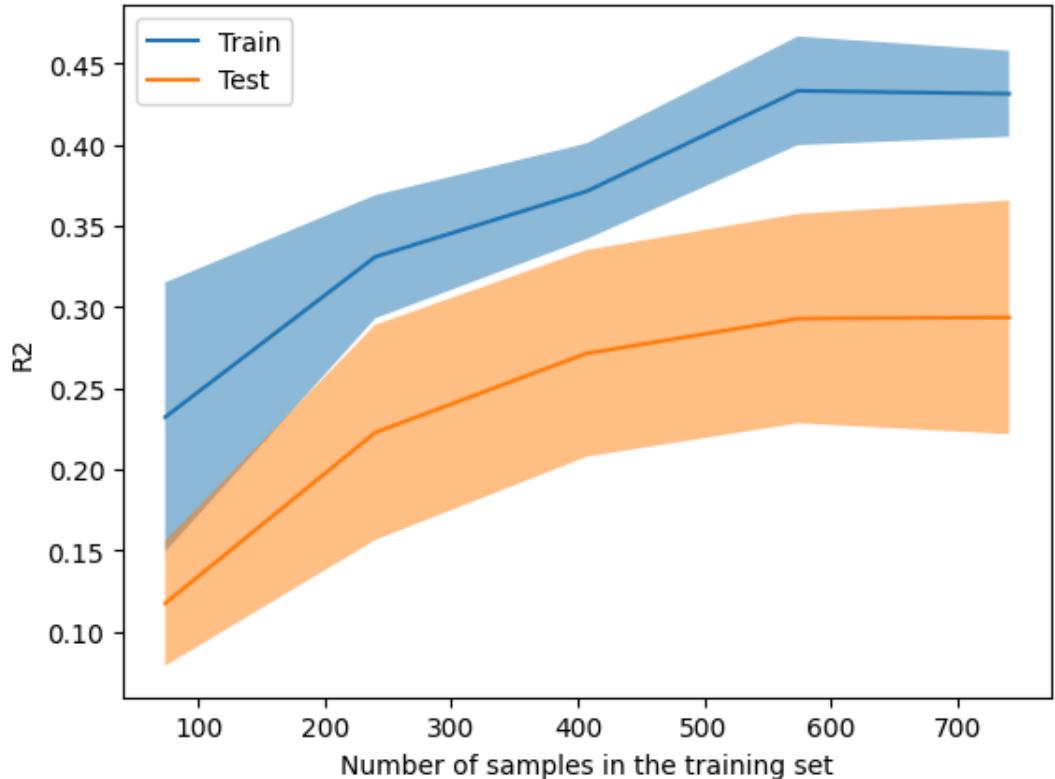
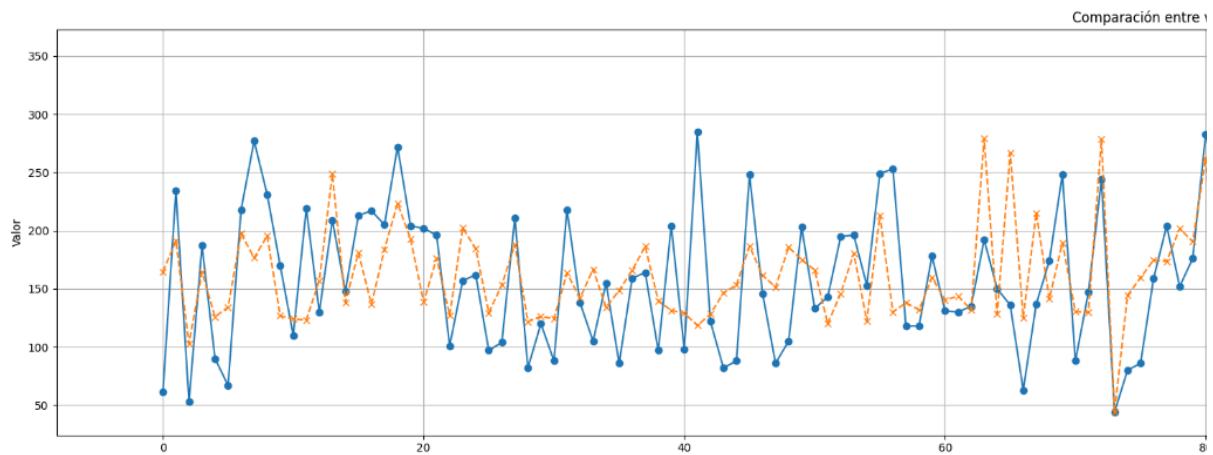


Figura 14: curvas de entrenamiento de TabPFN, experimento 1



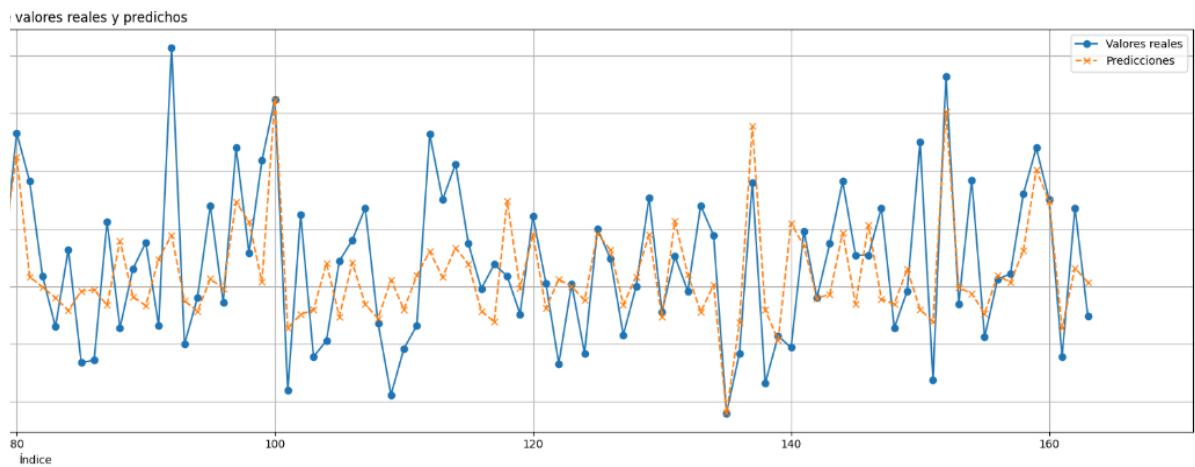


Figura 15: predicciones en el conjunto de evaluación para TabPFN, experimento 1

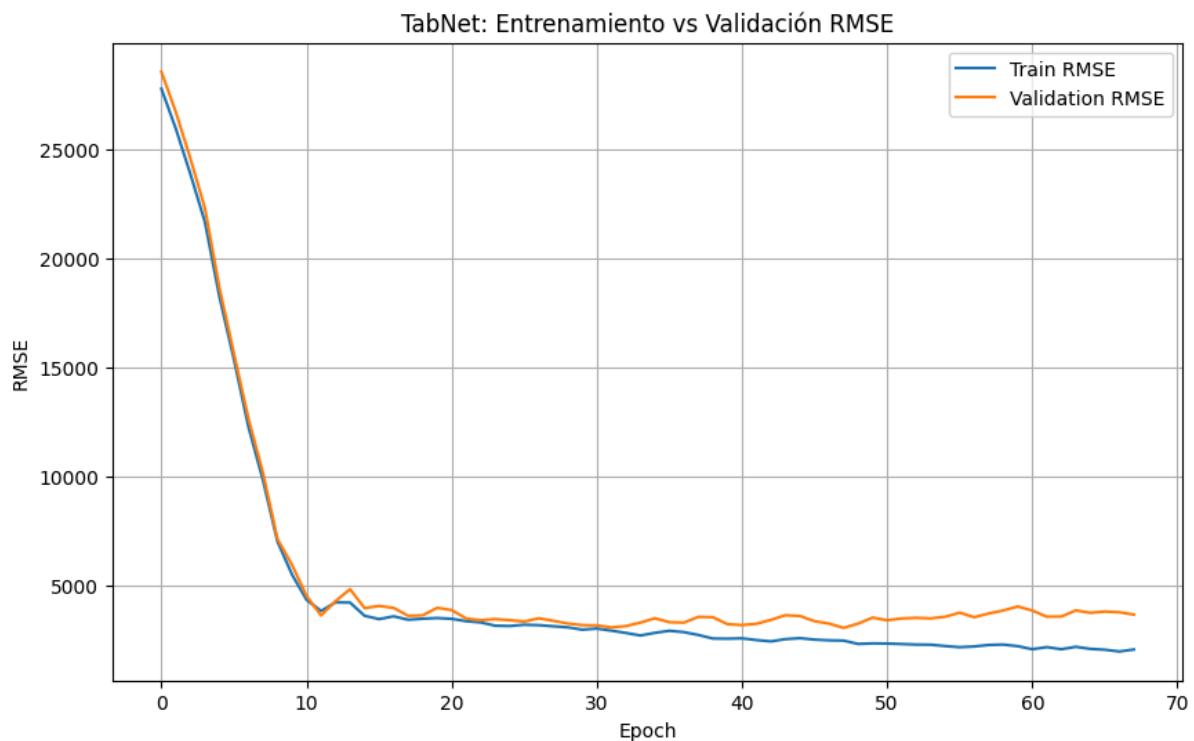


Figura 16: curvas de entrenamiento de TabNet, experimento 1

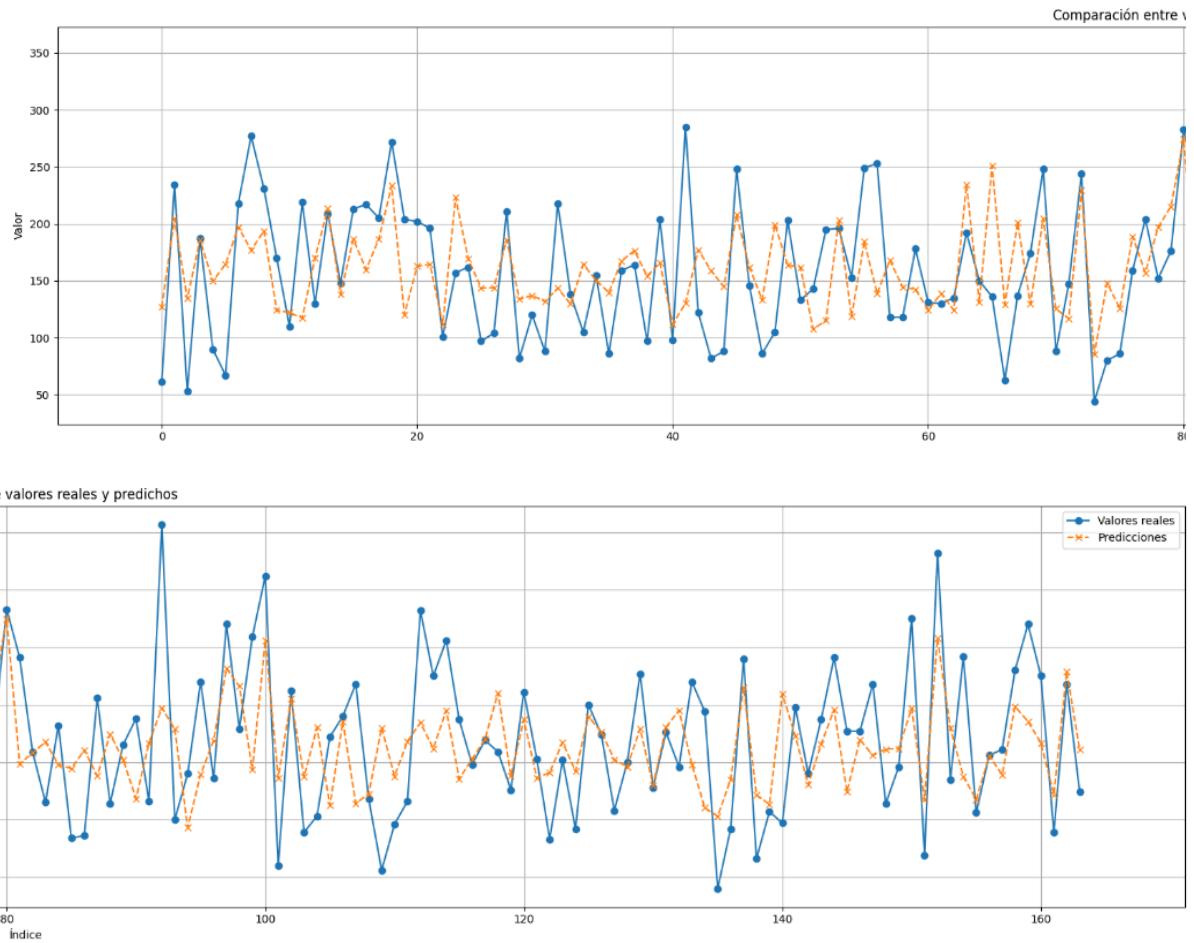


Figura 17: predicciones en el conjunto de evaluación para TabNet, experimento 1

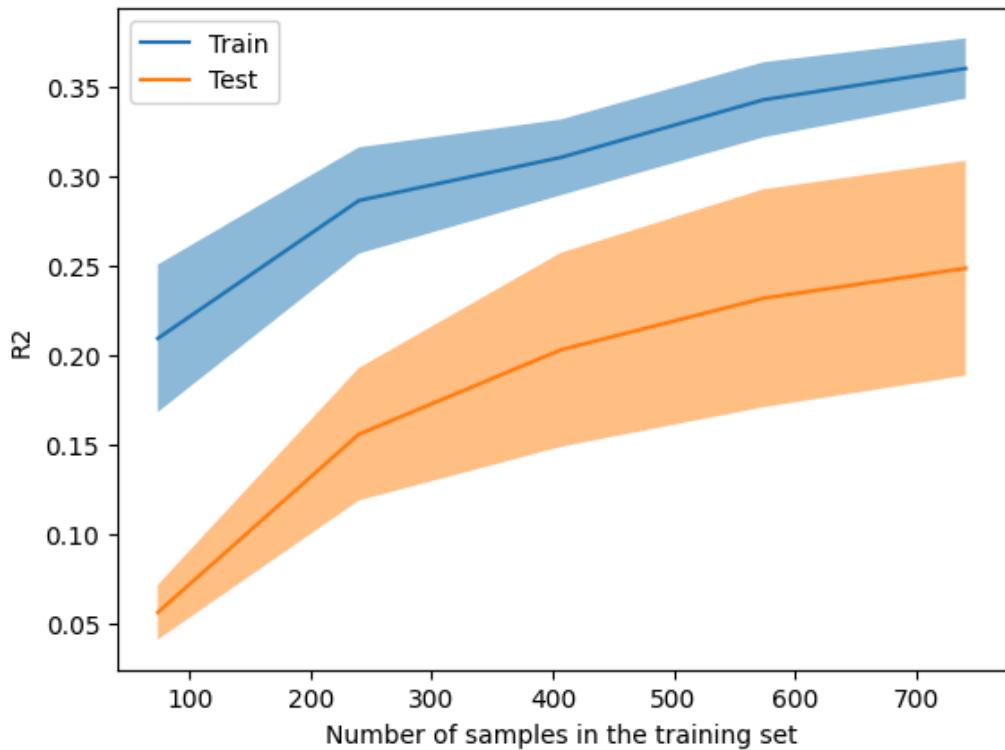


Figura 18: curvas de entrenamiento de SVM, experimento 1

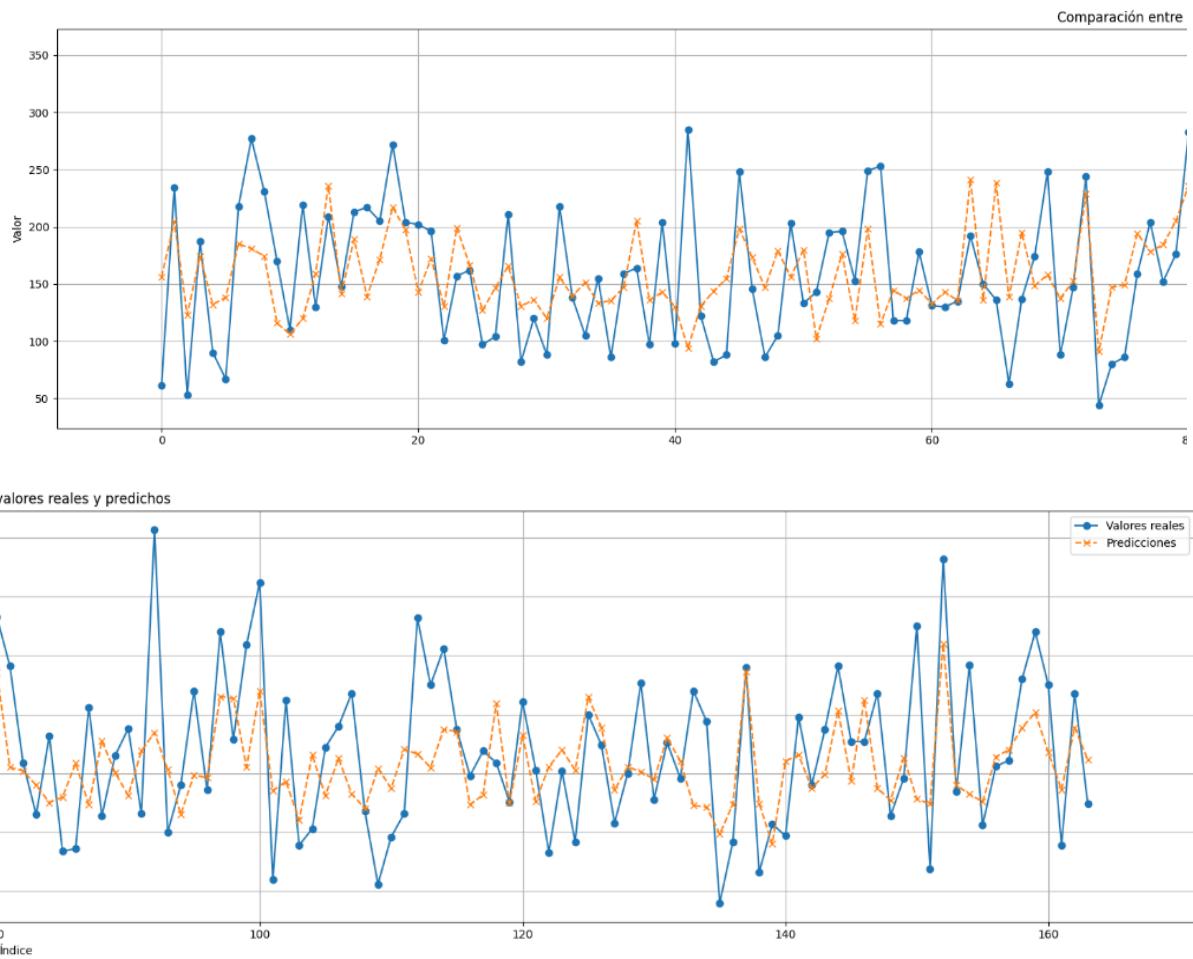


Figura 19: predicciones en el conjunto de evaluación para SVM, experimento 1

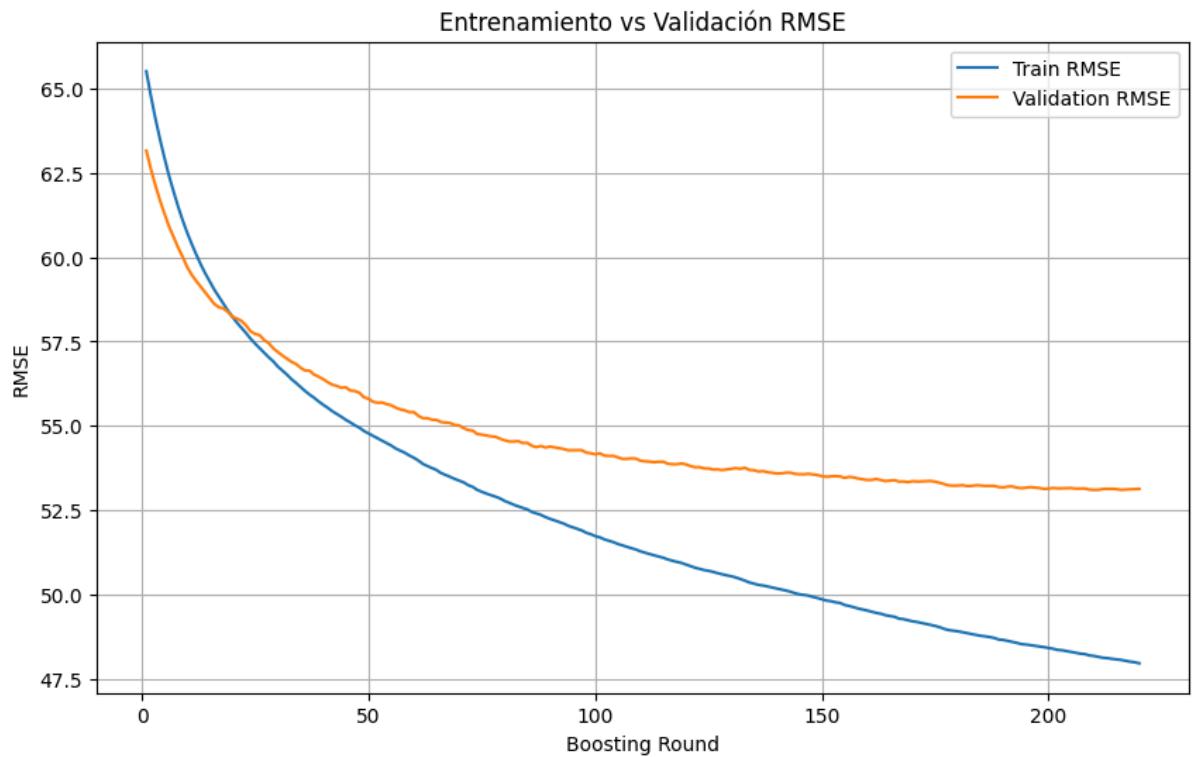


Figura 20: curvas de entrenamiento de XGBoost, experimento 1

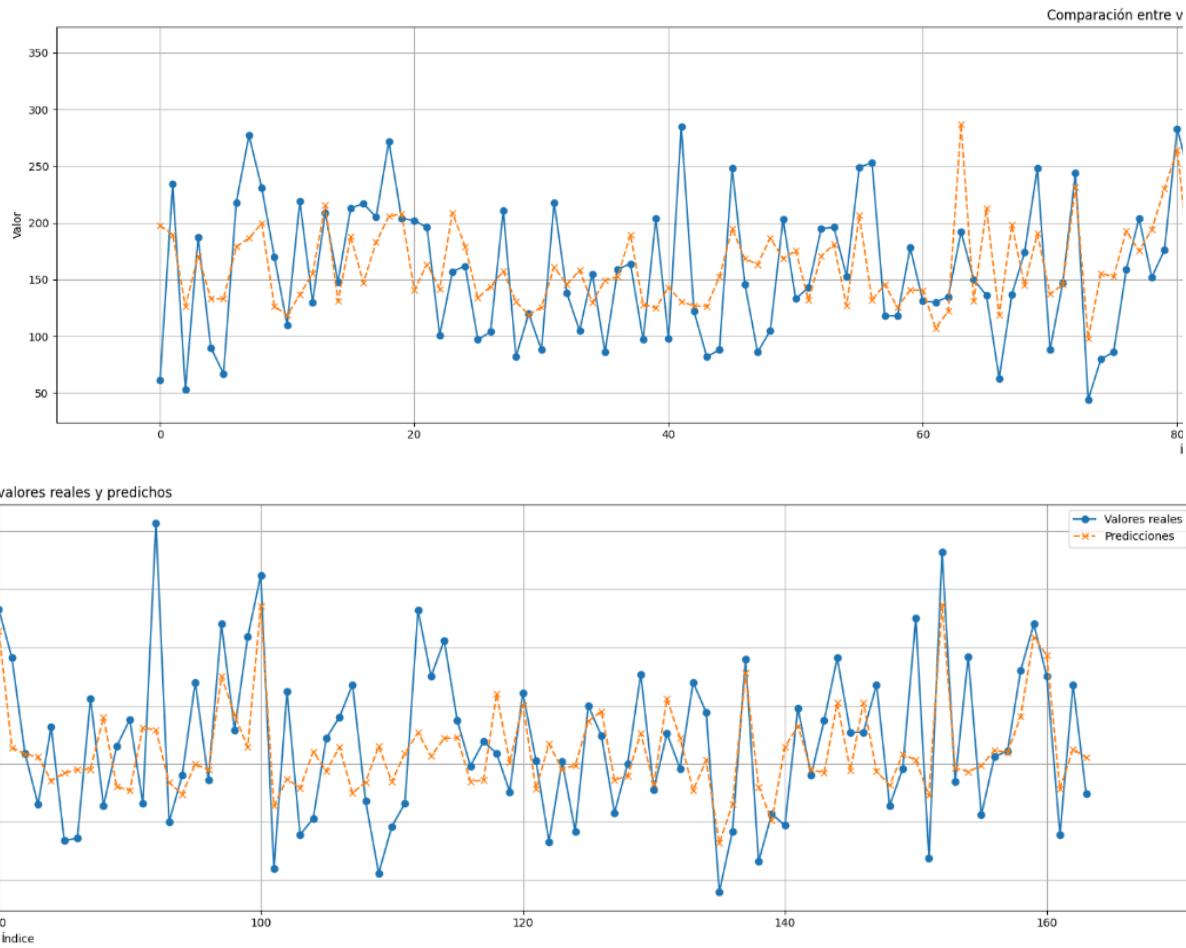


Figura 21: predicciones en el conjunto de evaluación para XGBoost, experimento 1

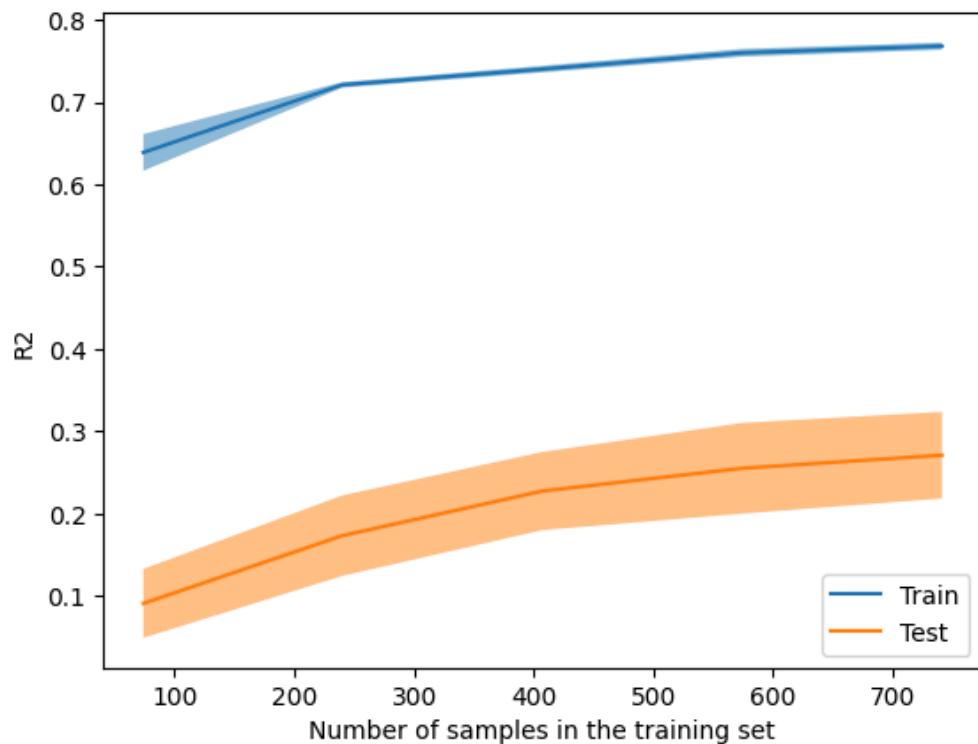


Figura 22: curvas de entrenamiento de Random Forest, experimento 1

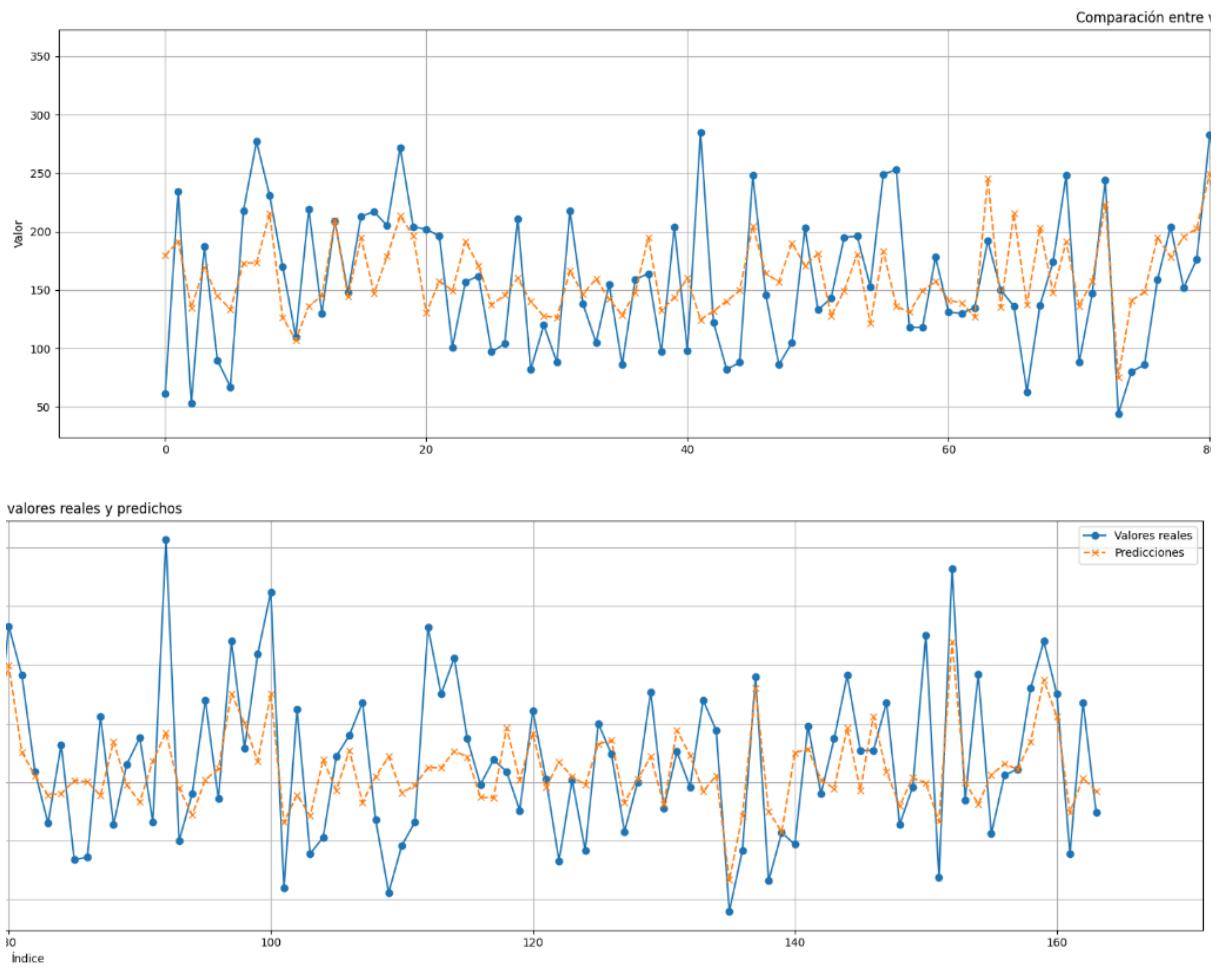


Figura 23: predicciones en el conjunto de evaluación para Random Forest, experimento 1

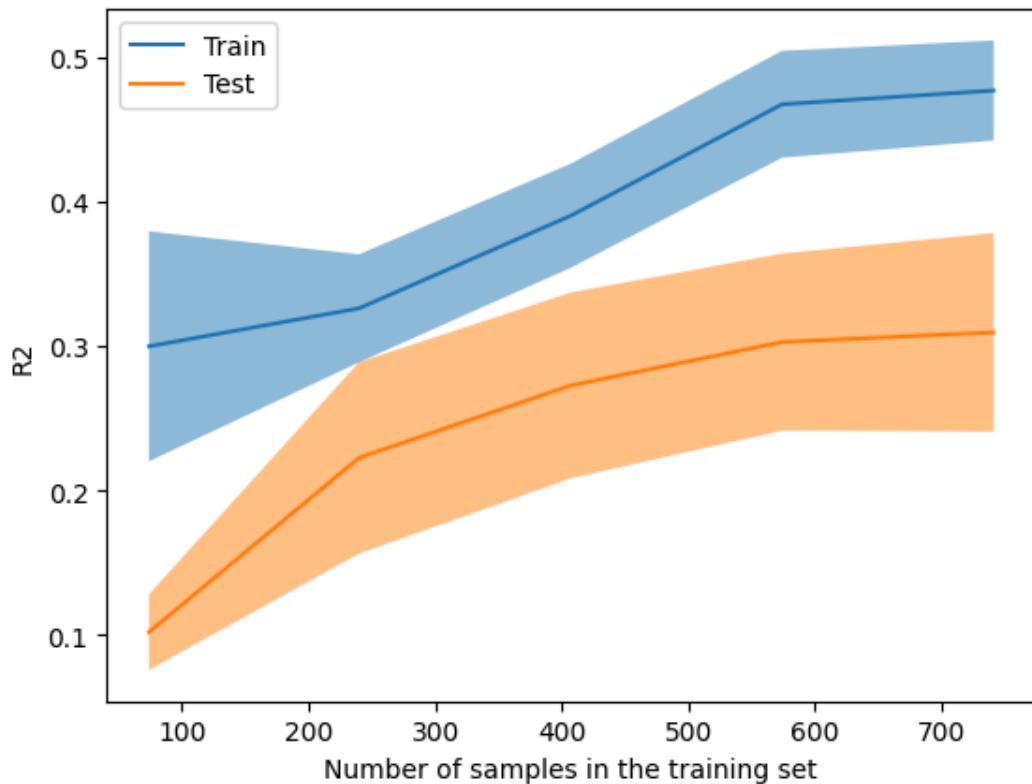


Figura 24: curvas de entrenamiento de TabPFN, experimento 2

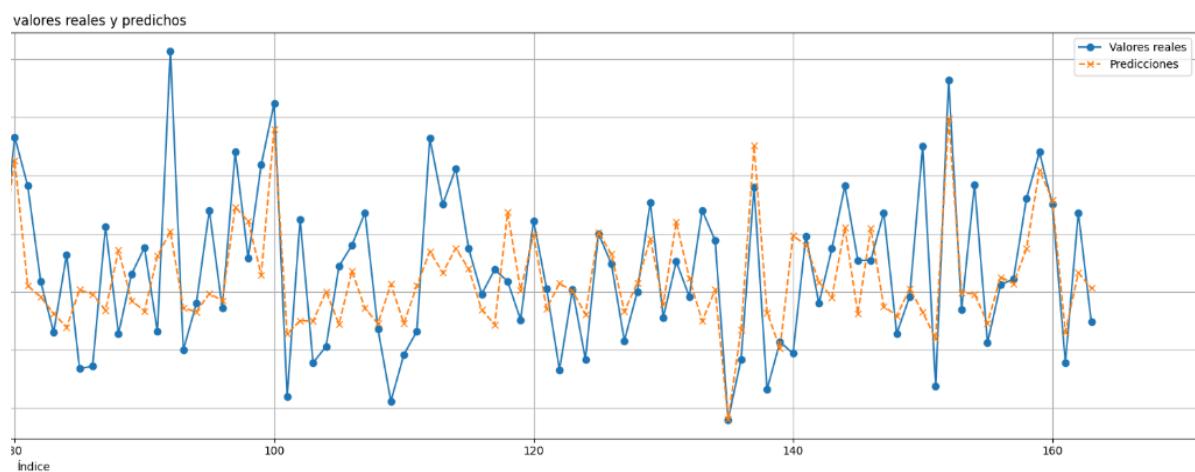
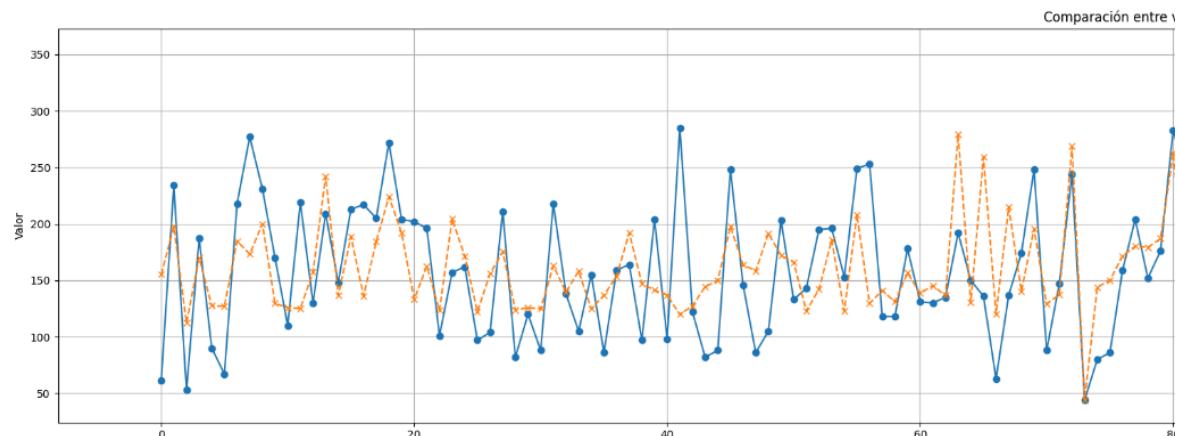


Figura 25: predicciones en el conjunto de evaluación para TabPFN, experimento 2

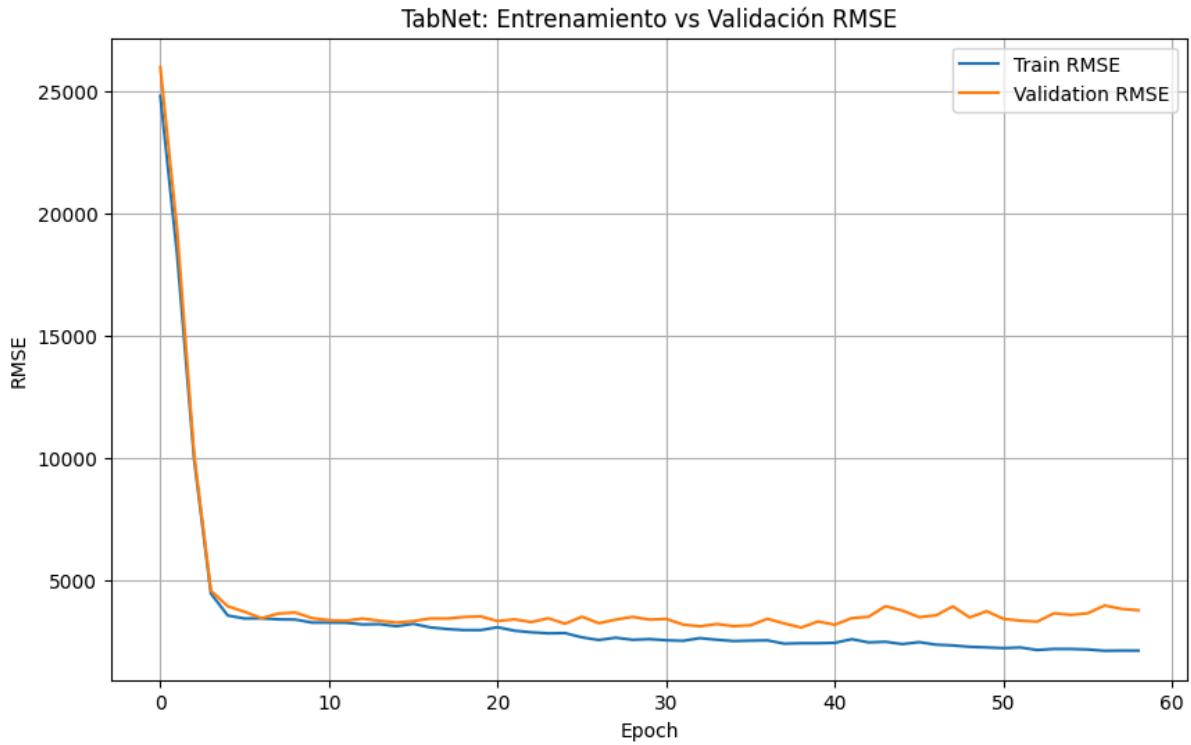


Figura 26: curvas de entrenamiento de TabNet, experimento 2

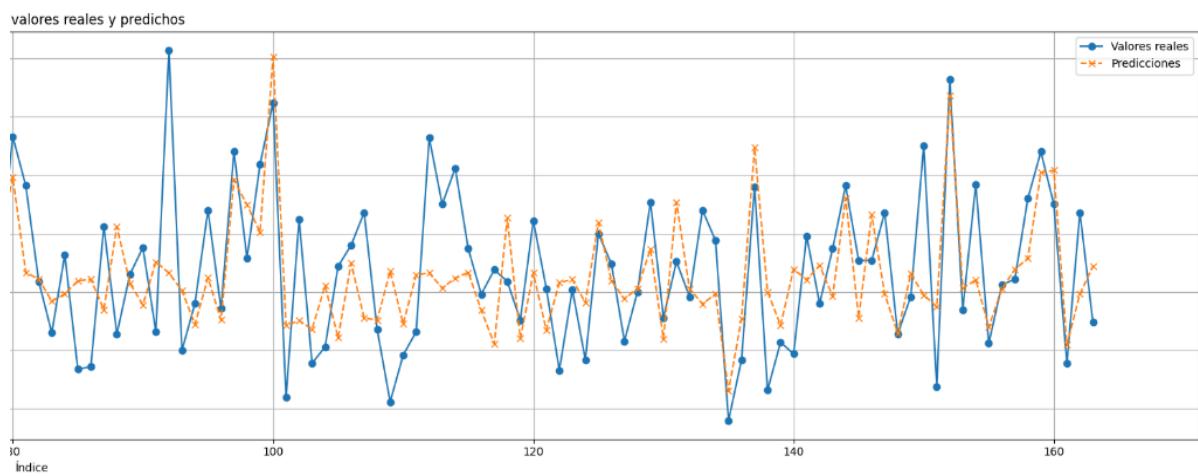
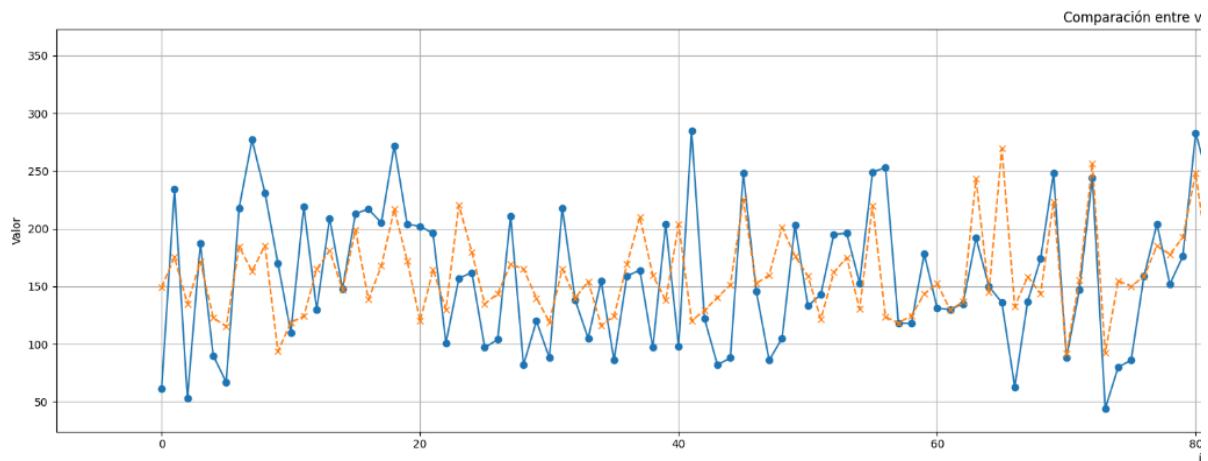


Figura 27: predicciones en el conjunto de evaluación para TabNet, experimento 2

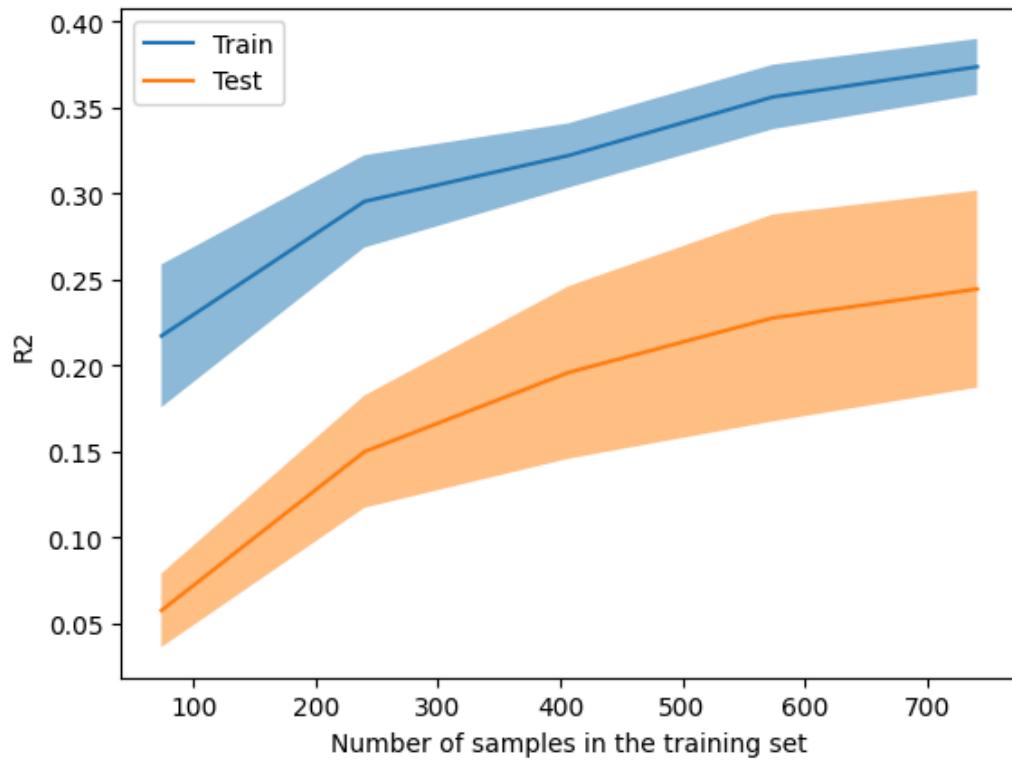


Figura 28: curvas de entrenamiento de SVM, experimento 2

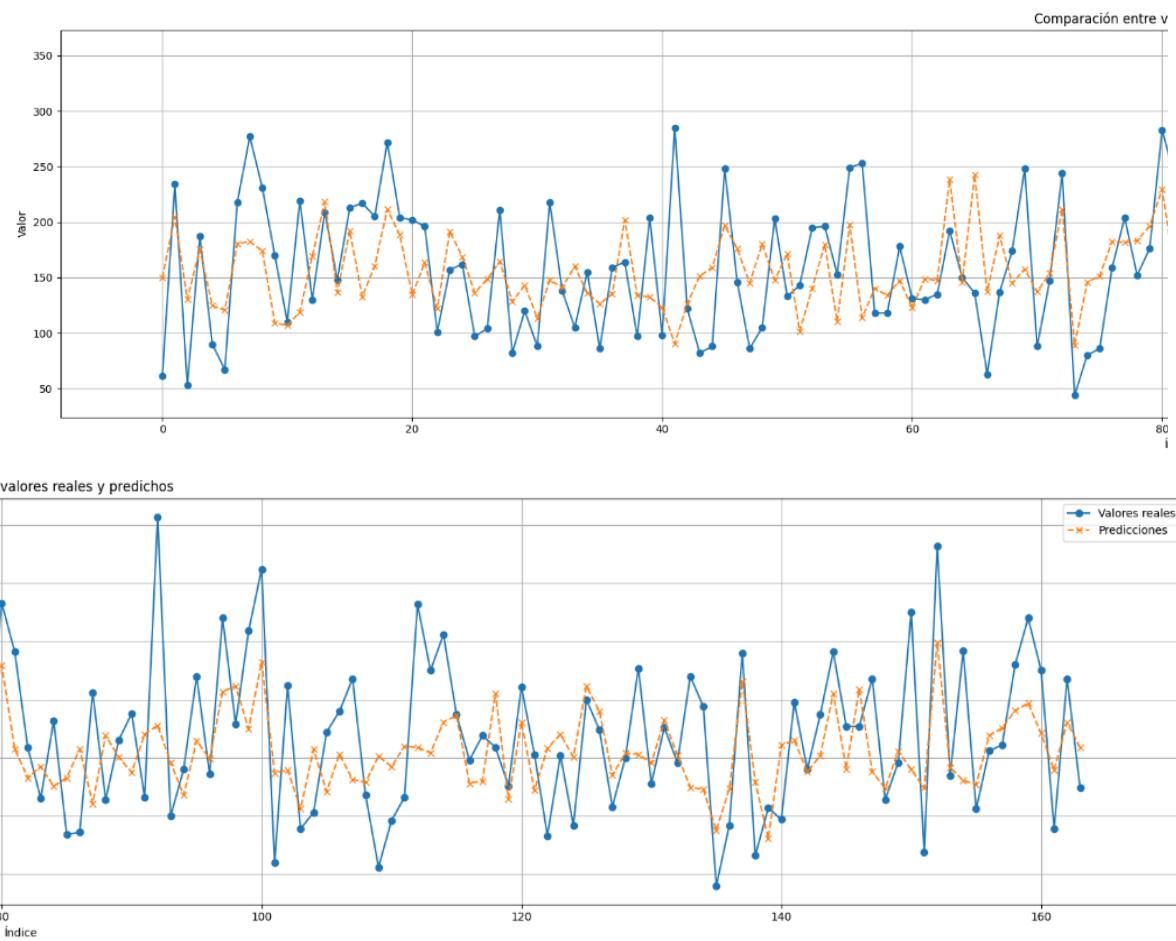


Figura 29: predicciones en el conjunto de evaluación para SVM, experimento 2

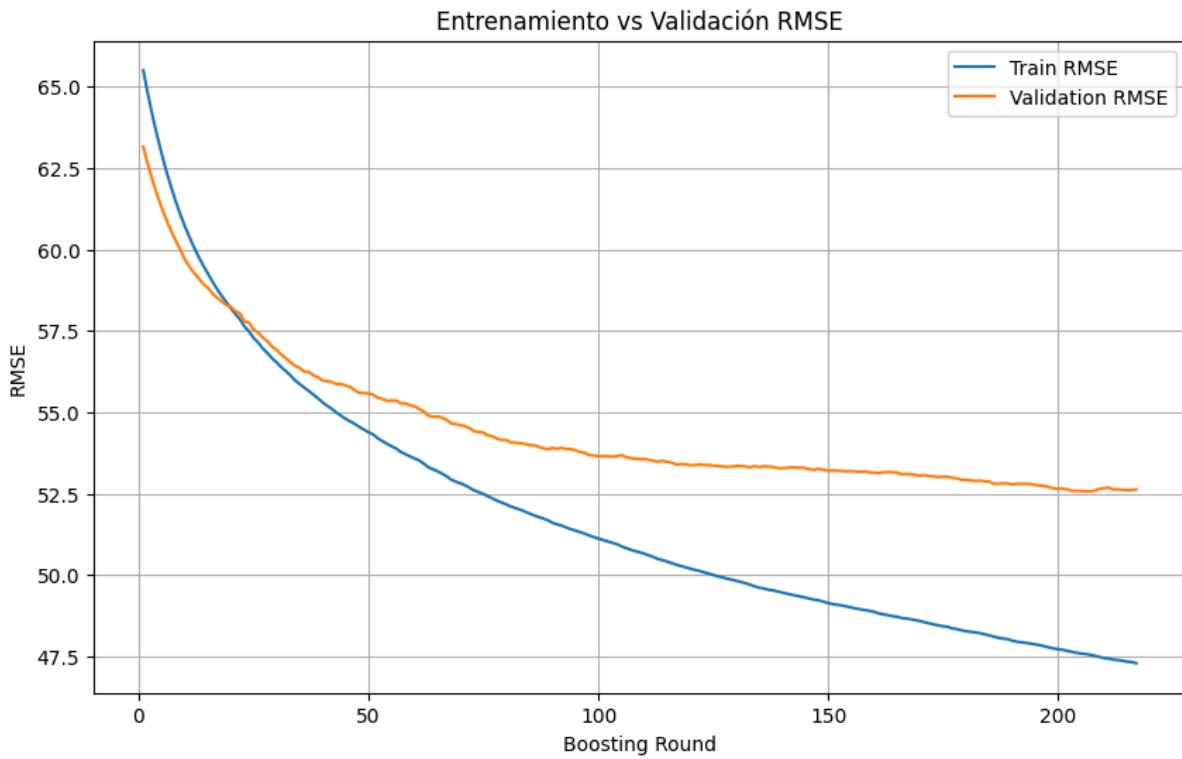


Figura 30: curvas de entrenamiento de XGBoost, experimento 2

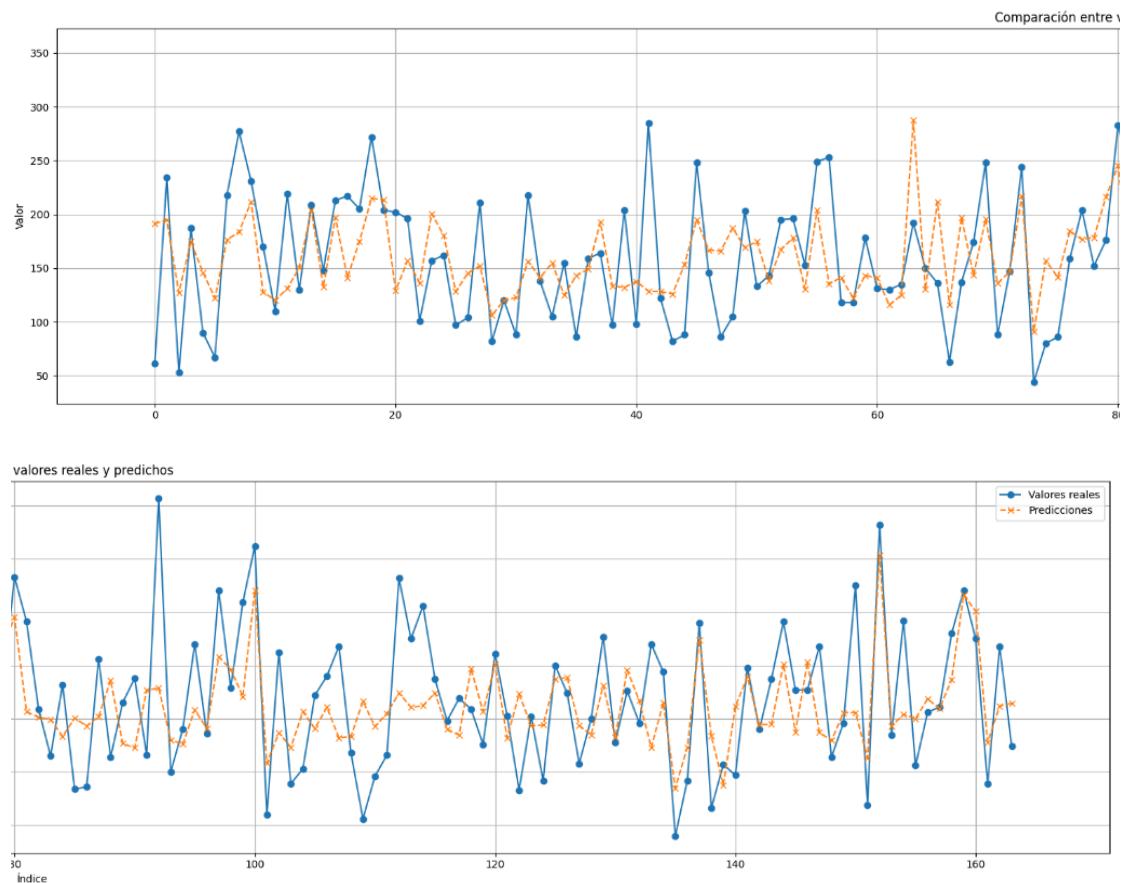


Figura 31: predicciones en el conjunto de evaluación para XGBoost, experimento 2

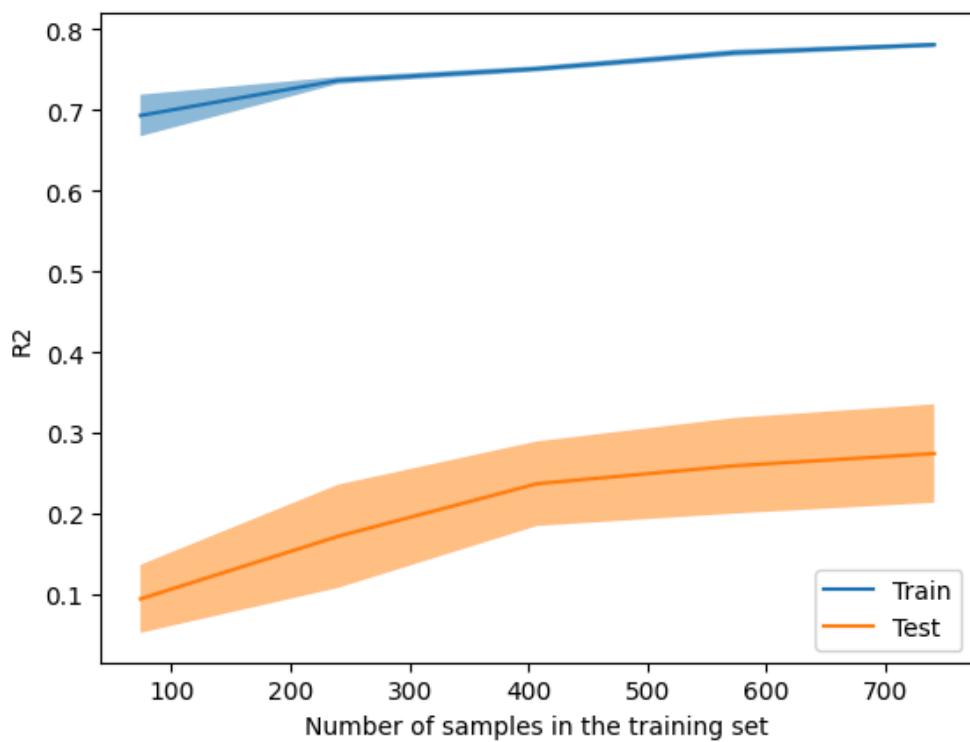


Figura 32: curvas de entrenamiento de Random Forest, experimento 2

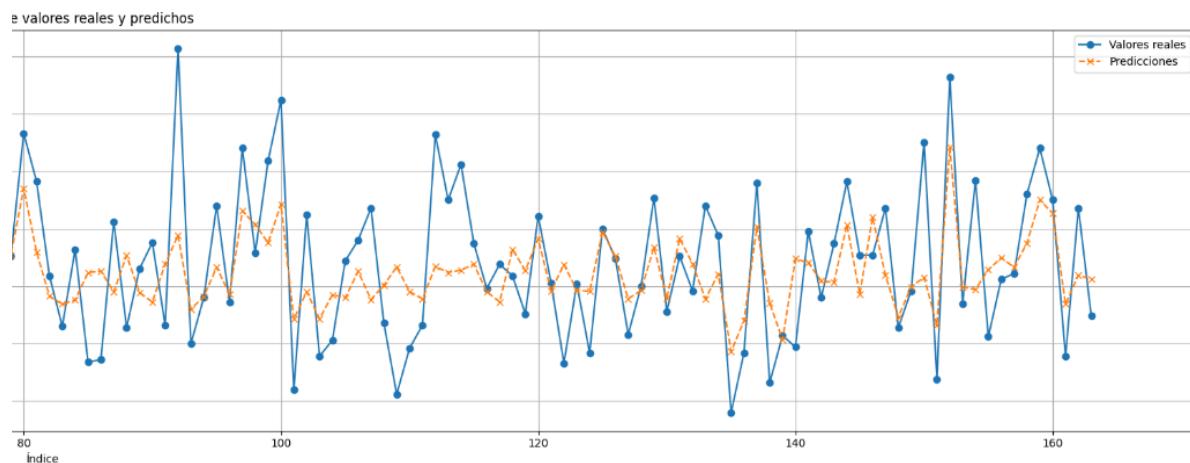
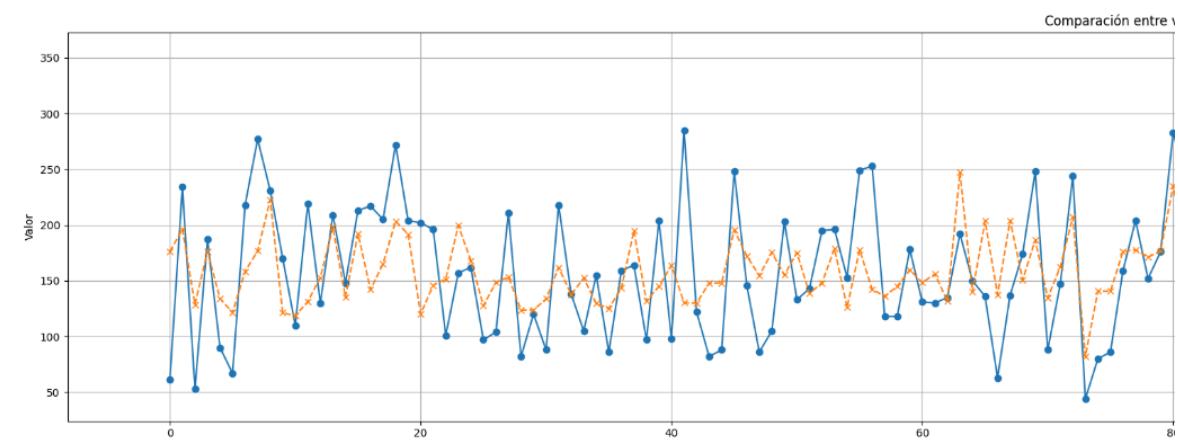


Figura 33: predicciones en el conjunto de evaluación para Random Forest, experimento 2

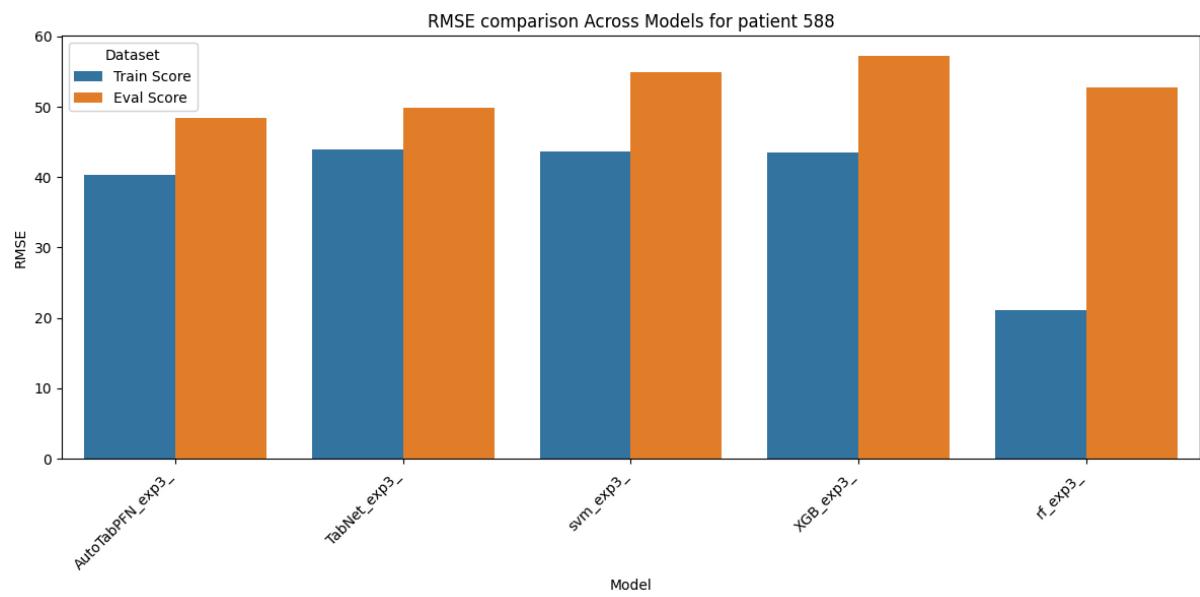


Figura 34: Comparativa de RMSE en entrenamiento y validación por modelo para el usuario 588

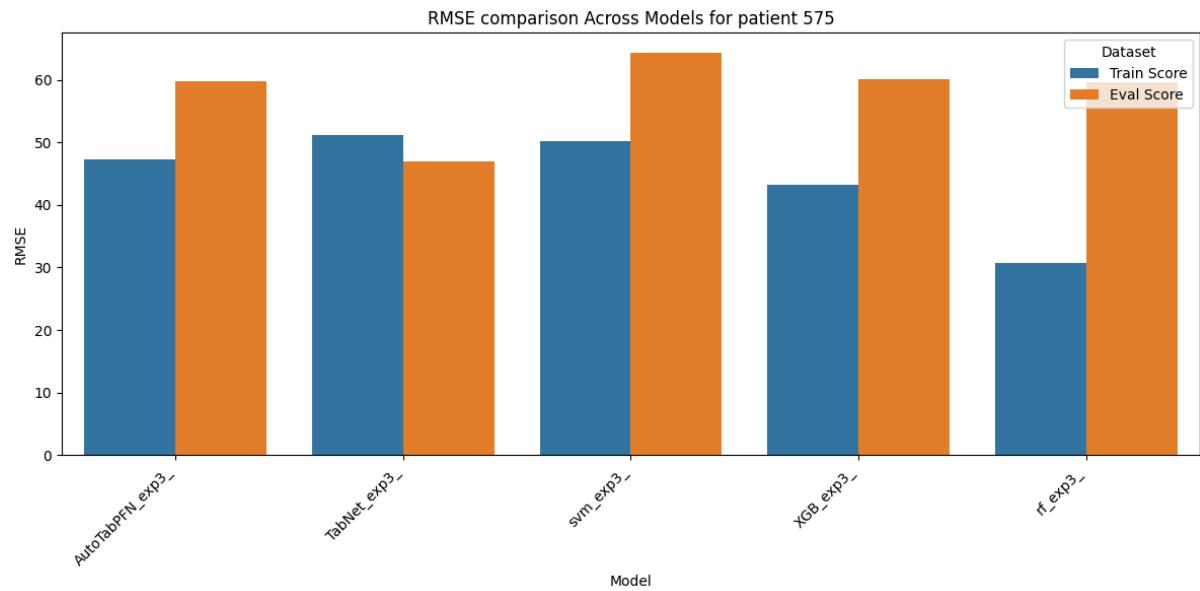


Figura 35: Comparativa de RMSE en entrenamiento y validación por modelo para el usuario 575

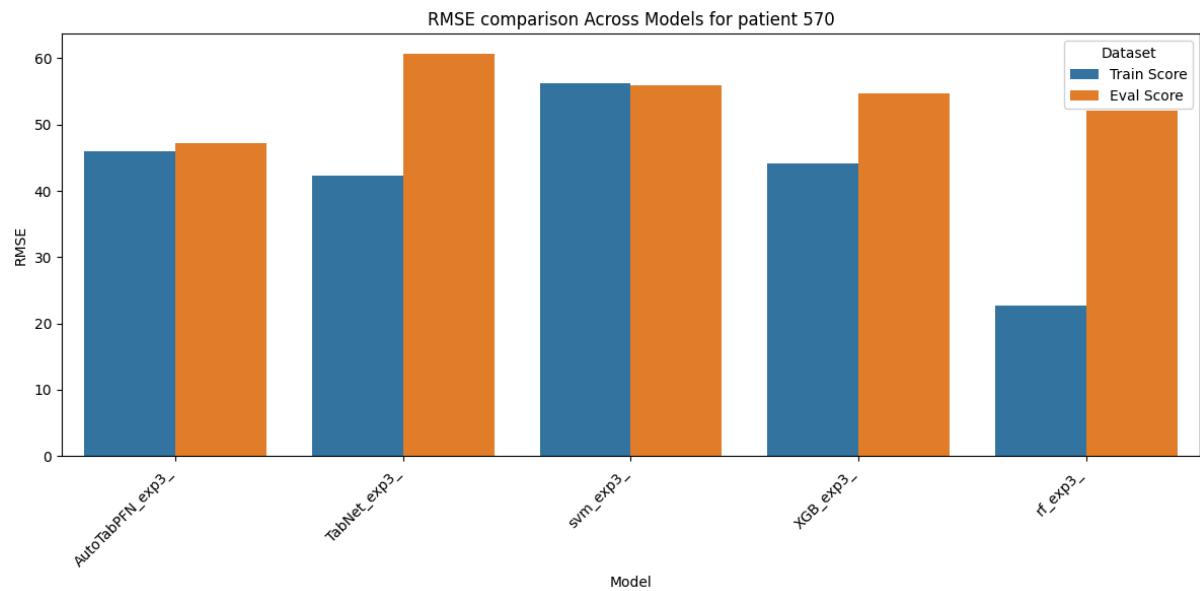


Figura 36: Comparativa de RMSE en entrenamiento y validación por modelo para el usuario 570

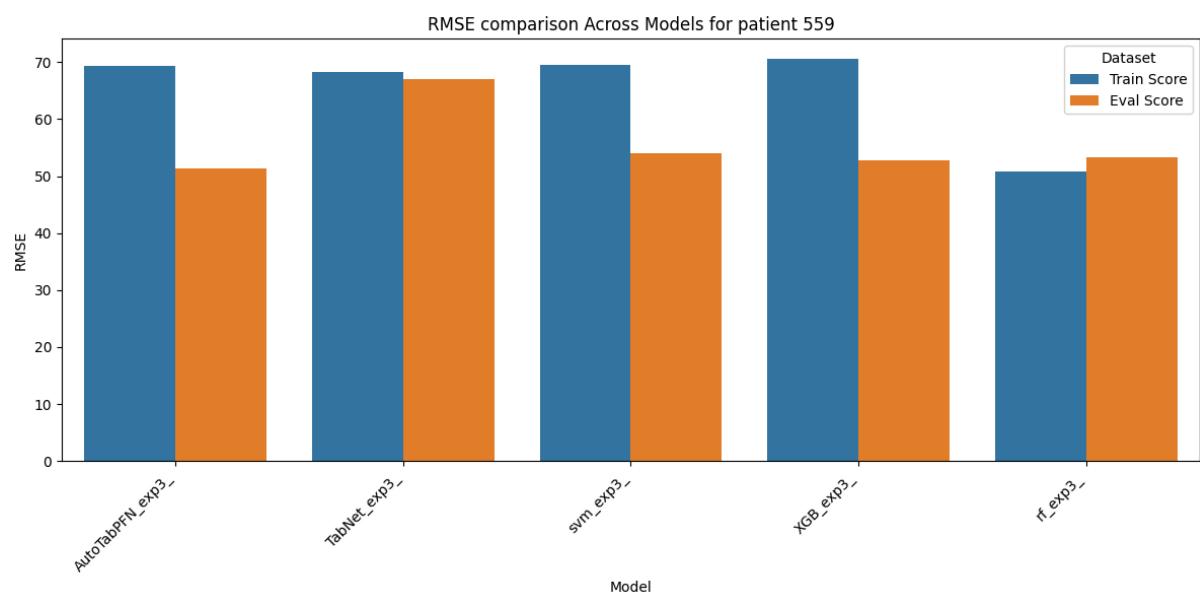


Figura 37: Comparativa de RMSE en entrenamiento y validación por modelo para el usuario 559

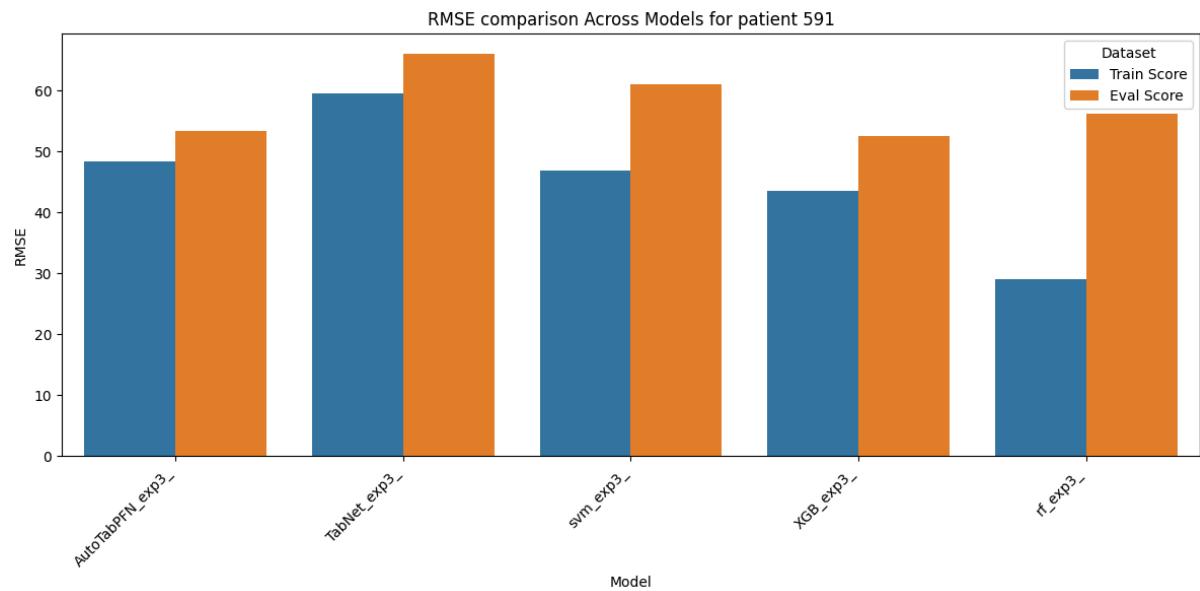


Figura 38: Comparativa de RMSE en entrenamiento y validación por modelo para el usuario 591

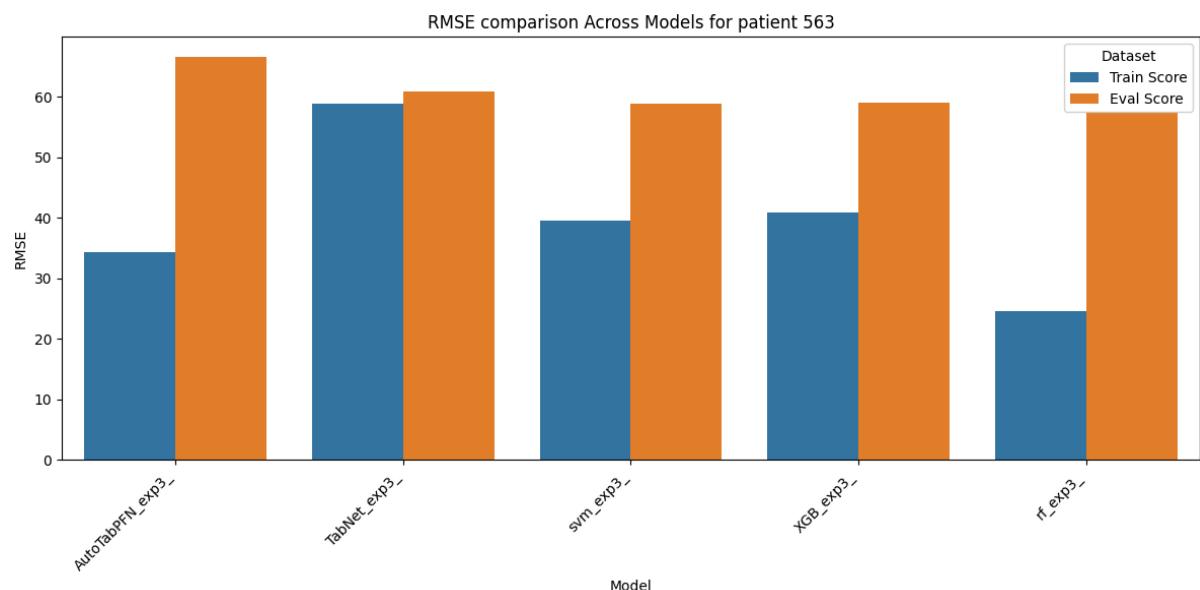


Figura 39: Comparativa de RMSE en entrenamiento y validación por modelo para el usuario 563

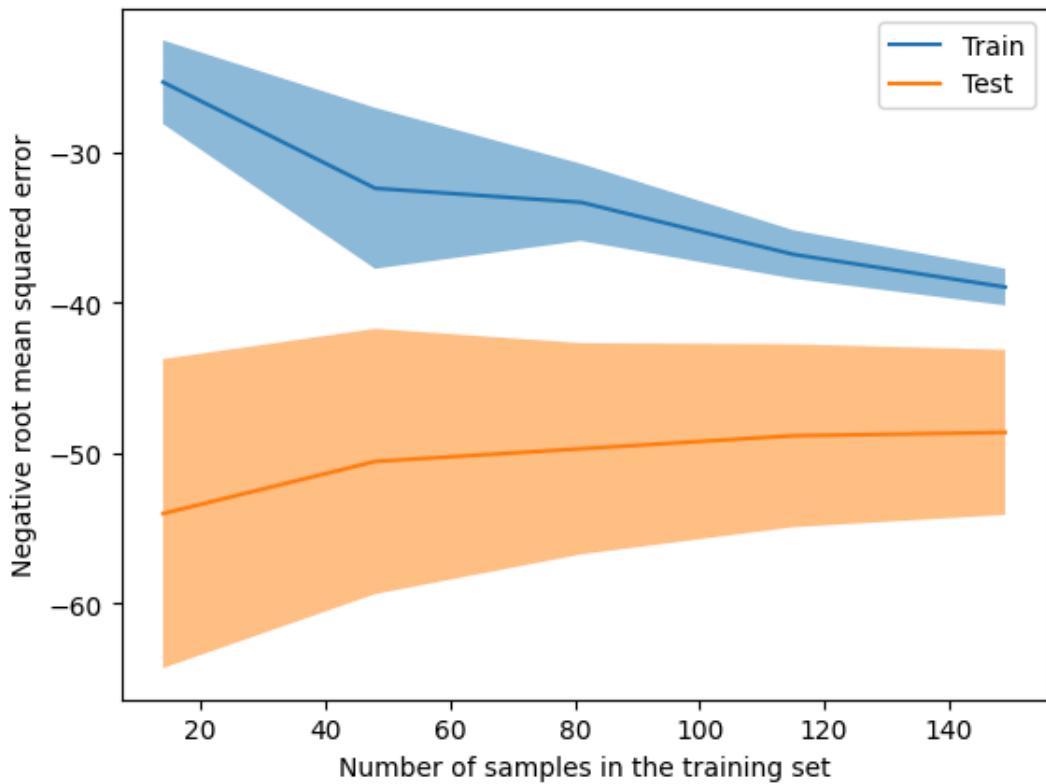


Figura 40: Curvas de entrenamiento para TabPFN, experimento 3

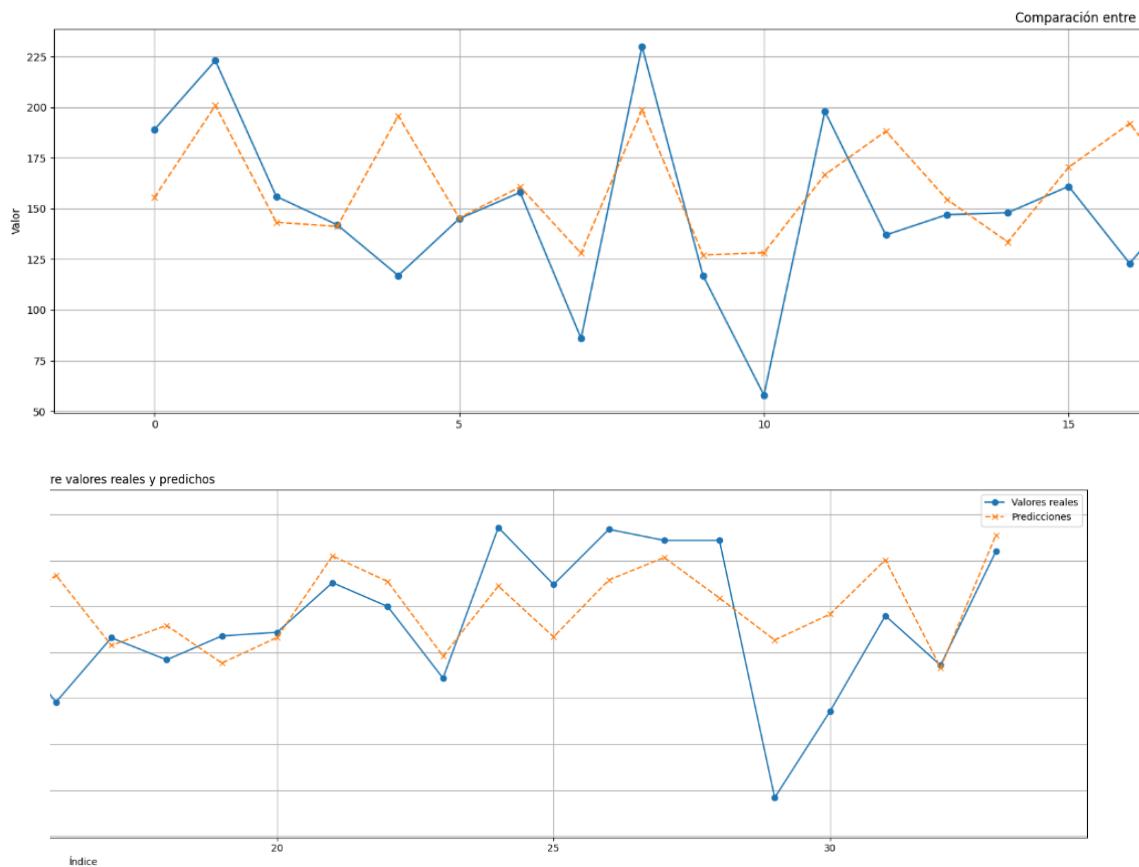


Figura 41: Muestra de predicciones de evaluación TabPFN, experimento 3

TabNet: Entrenamiento vs Validación RMSE

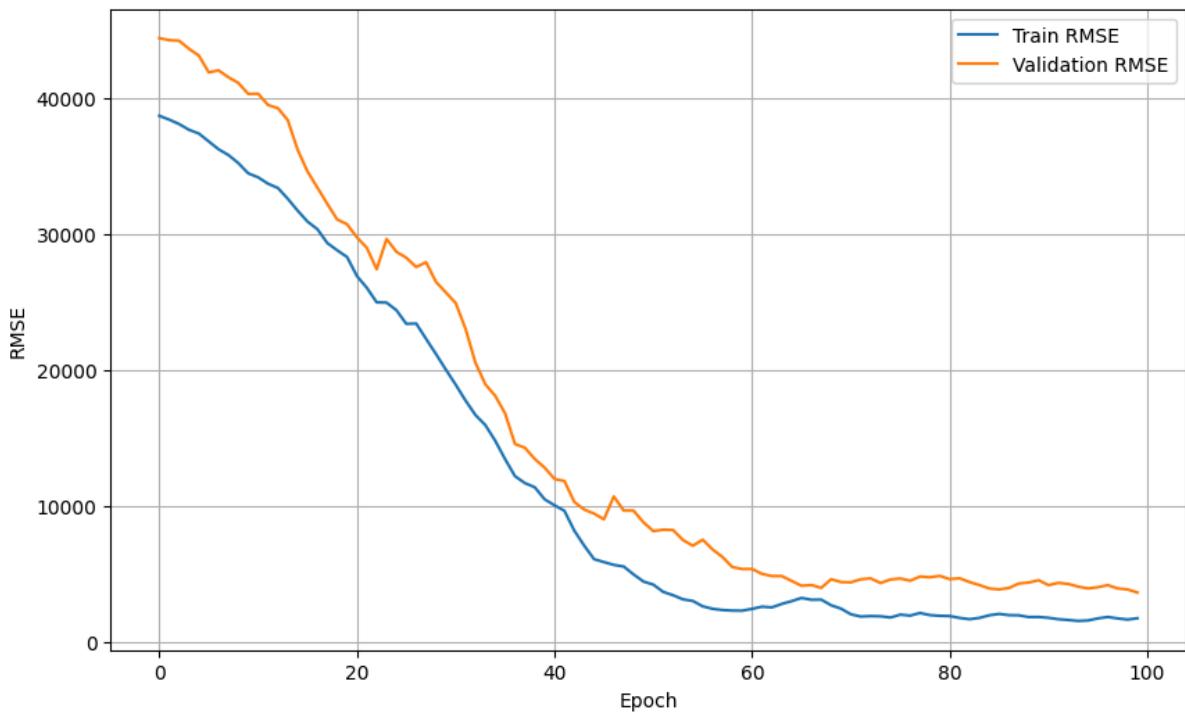


Figura 42: Muestra de curvas de entrenamiento TabNet, experimento 3

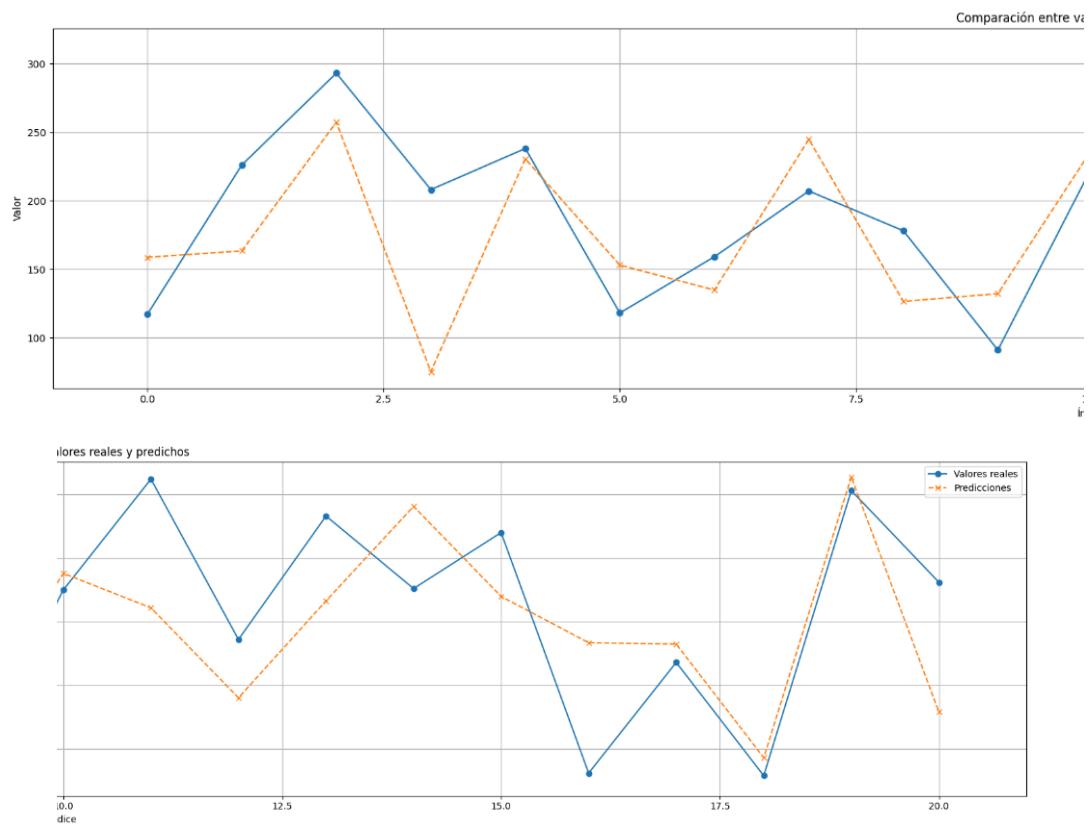


Figura 43: Muestra de predicciones en evaluación para TabNet, experimento 3

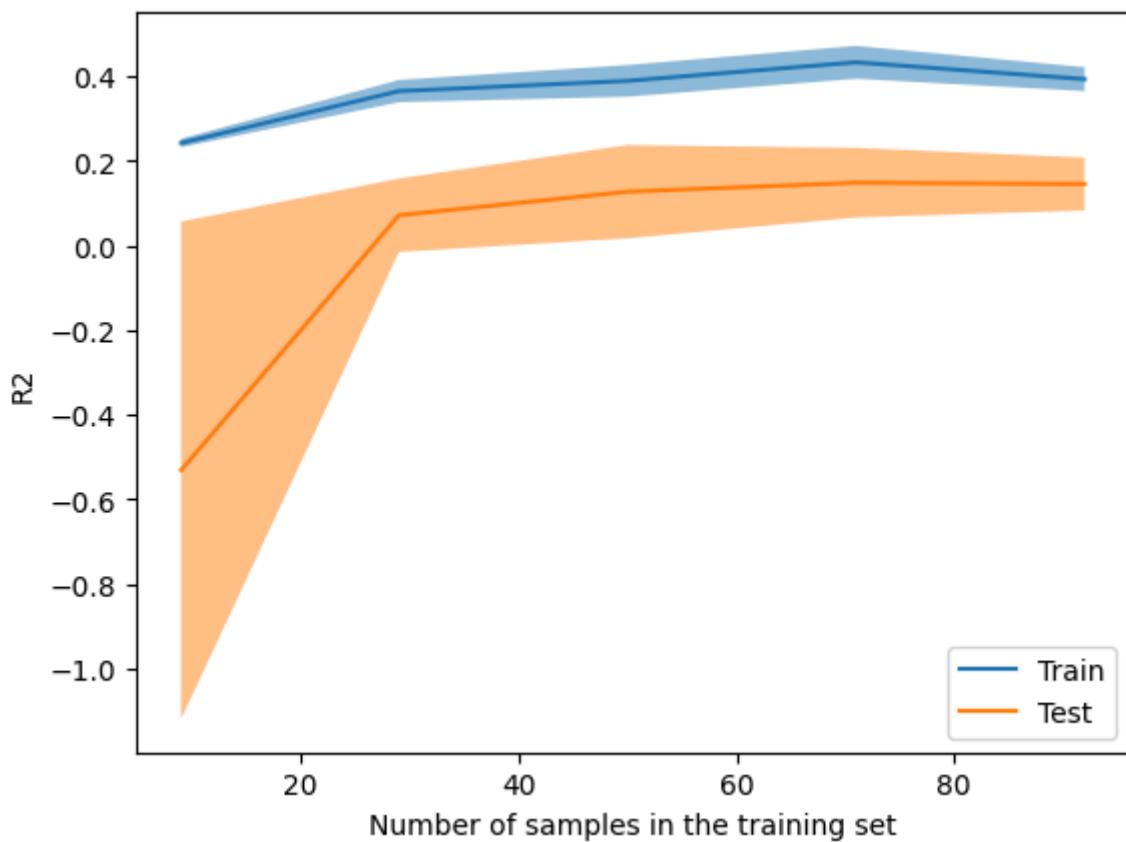


Figura 44: Muestra de curvas de entrenamiento para SVM, experimento 3

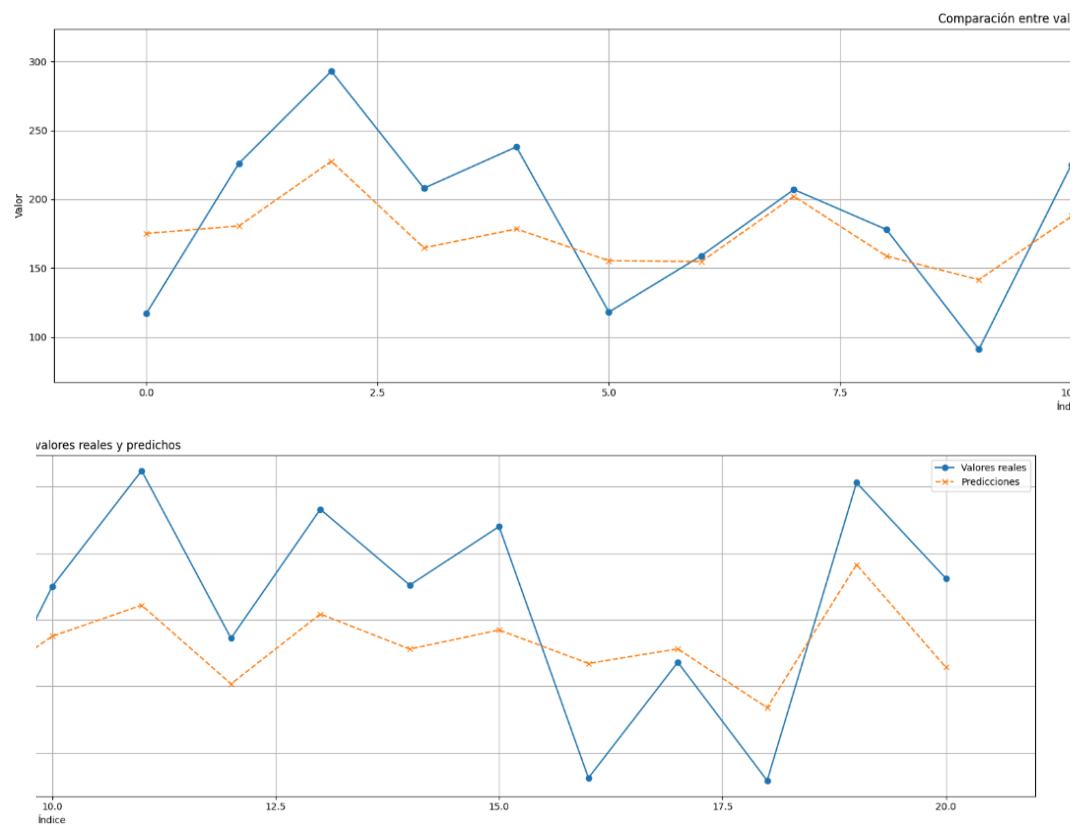


Figura 45: Muestra de predicciones en evaluación para SVM, experimento 3

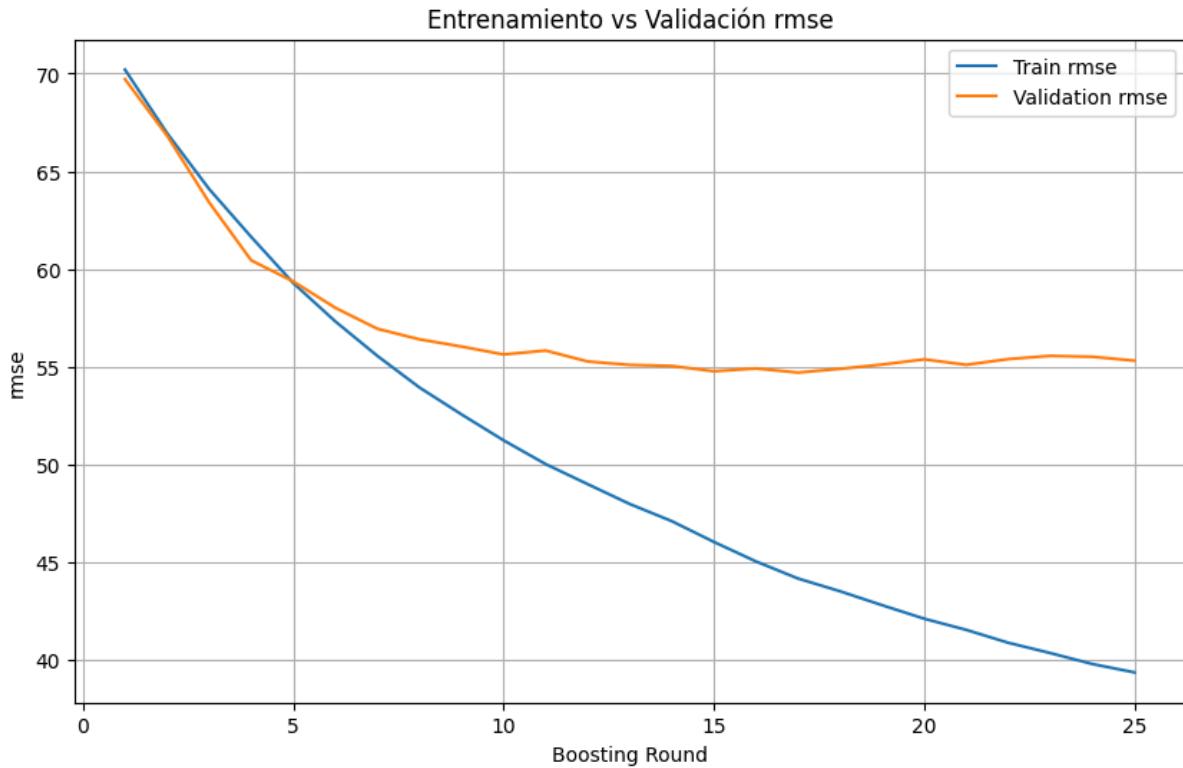


Figura 46: Muestra de curvas de entrenamiento para XGBoost, experimento 3

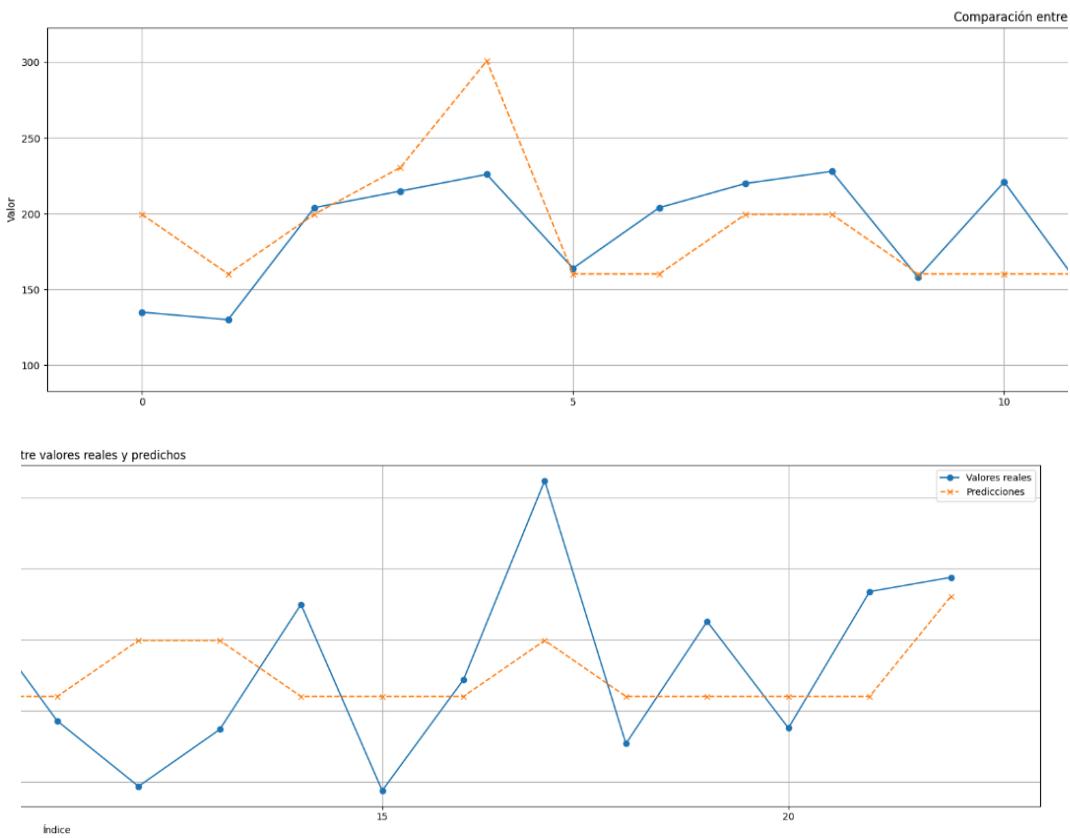


Figura 47: Muestra de predicciones en evaluación para XGBoost, experimento 3

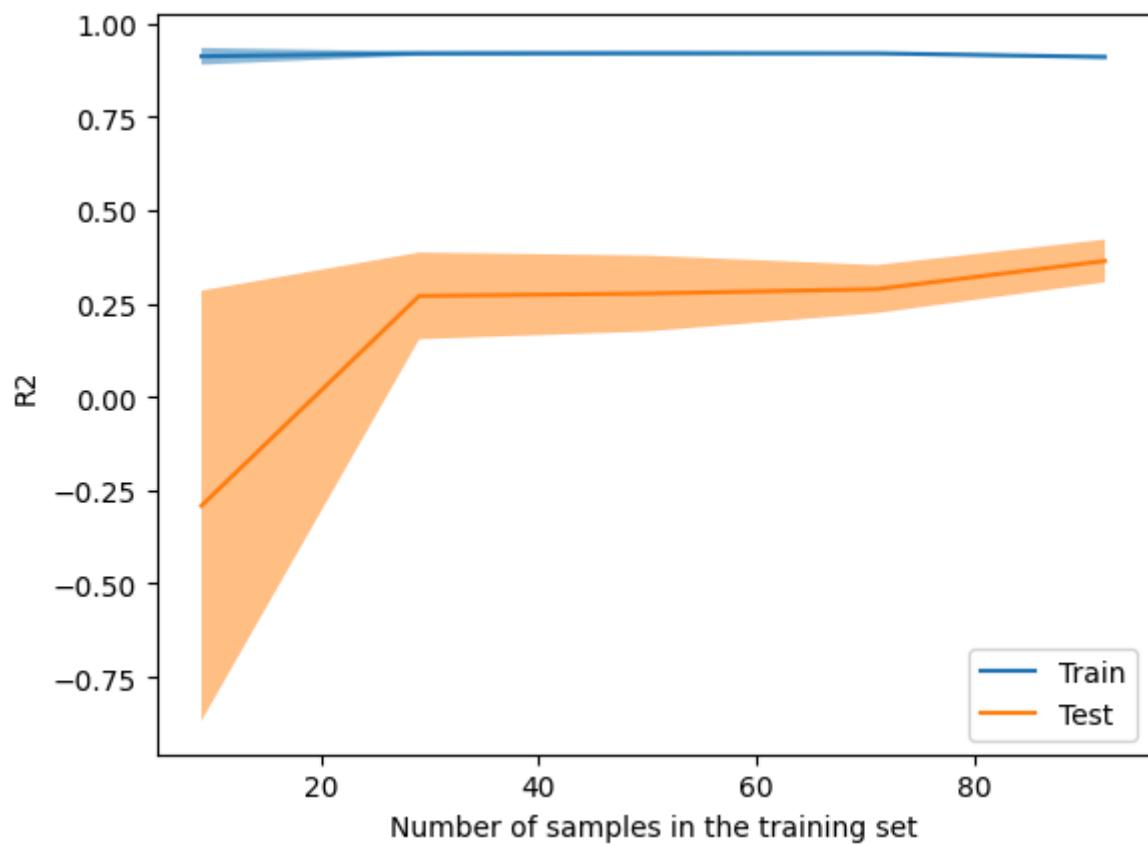


Figura 48: Muestra de curvas de entrenamiento para Random Forest, experimento 3

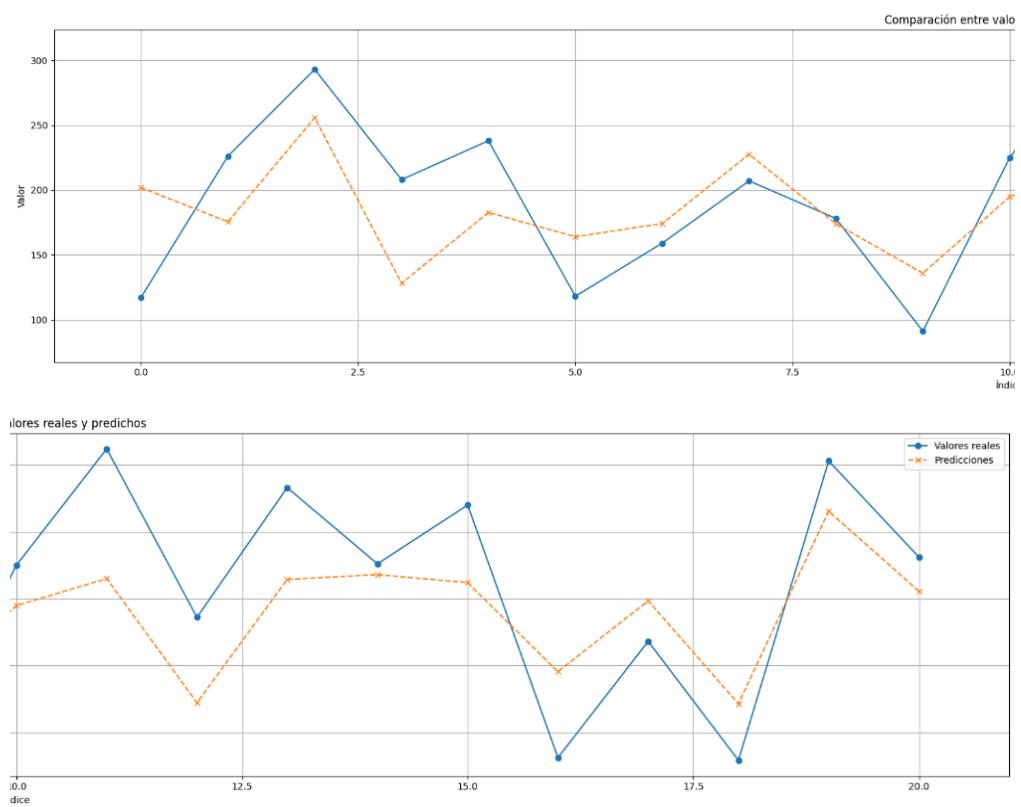


Figura 49: Muestra de predicciones en evaluación para Random Forest, experimento 3