Abstract

El objetivo principal de este estudio es estimar la incidencia de diabetes en un conjunto de pacientes de la Obra Social OSEP de la región de Cuyo. Para ello se desarrollan dos modelos de regresión para predecir el nivel de glucemia en ayunas en pacientes con riesgo de desarrollar diabetes tipo 2. La ausencia de este valor en las encuestas y en las bases de datos de laboratorio limita la capacidad de los profesionales de la salud para tomar decisiones adecuadas y precisas relativas a la prevención y al tratamiento de la diabetes tipo 2.

Este estudio presenta dos modelos de regresión que permiten estimar el valor de glucemia en ayunas, que pretenden contribuir significativamente en la detección temprana de esta enfermedad en pacientes en los que este valor no se encuentra disponible. Los modelos se construyeron utilizando datos de laboratorio y de la encuesta Findrisc, y se evaluaron mediante validación cruzada y otras métricas de desempeño.

Asimismo, se espera que este estudio pueda contribuir con la práctica clínica, proporcionando a los profesionales de la salud una herramienta valiosa para mejorar la precisión del diagnóstico de diabetes tipo 2 y mejorar la calidad de vida de los pacientes afectados por esta enfermedad ayudándolos en su detección temprana.

Palabras claves

Diabetes tipo II, Findrisc, Machine learning, Regresión, Ingeniería de Atributos, Glucemia.

Índice general

1. Capítulo 1: Introducción

1.1. Motivación e importancia del campo

La Diabetes Mellitus Tipo 2 (DM2) es un trastorno endocrino-metabólico crónico. Resulta de una alteración metabólica de los glúcidos y lípidos originada a partir de una compleja interacción entre factores de tipo genético y variables ambientales. Algunos de los mecanismos que contribuyen a la hiperglucemia son la deficiencia en la secreción de insulina, la disminución en el uso de la glucosa o el aumento de su producción. Esta enfermedad genera un deterioro progresivo y silencioso de ciertos órganos que deviene en secuelas de carácter permanente y en los casos más graves puede ocasionar la muerte. La DM2 puede causar la ceguera, la enfermedad renal terminal, la amputación traumática y la enfermedad cardiovascular. En casi todas las sociedades desarrolladas la diabetes es una de las principales causas de ceguera, amputaciones y enfermedad renal terminal (Ruiz de Adana, 2012). La esperanza de vida de los pacientes diabéticos es menor que la de la población general.

El desarrollo de la DM2 está condicionado por factores genéticos y ambientales, aceptándose que la enfermedad sólo se desarrolla con mayor probabilidad en personas con predisposición genética. Si un padre es diabético la probabilidad de desarrollar la enfermedad es del 40% mientras que si los dos padres lo son la probabilidad asciende al 70%. Esta enfermedad favorece el desarrollo de factores de riesgo cardiovascular (FRCV) que pueden progresar a complicaciones crónicas que son responsables de su alta morbimortalidad, costo económico y deterioro de la calidad de vida (Gagliardino, 2016).

Las características descritas hacen de la DM2 una enfermedad con un importante impacto tanto a nivel individual, familiar y en el sistema sanitario en su conjunto. Las comorbilidades generan un costo de alrededor de 1.427 USD por persona (International Diabetes Federation, 2013).

Desde la perspectiva sanitaria la diabetes se ha convertido en uno de los problemas más graves de los últimos tiempos. Ha alcanzado proporciones de características epidémicas en la mayor parte del mundo (Ruiz de Adana, 2012). Se estima que existen actualmente 285 millones de personas afectadas a lo ancho del planeta, y esta cifra aumentará en los próximos años, alcanzando 438 millones en el año 2030 si se cumplen las últimas predicciones. La mayoría de los casos corresponden a diabetes mellitus tipo 2(DM2) (International Diabetes Federation, 2009).

La prevalencia a nivel mundial de la DM2 tiene una alta tasa de crecimiento debido probablemente a cambios en los hábitos alimentarios, a la disminución de la actividad física y el aumento de población con conductas sedentarias (Feskens, y otros, 1995).

Dado que se ha reportado que estratos socioculturales más bajos están más expuestos al desarrollo de esta patología, es una excelente oportunidad para el desarrollo de políticas nacionales en salud, que incorporen instancias formativas y participativas para la prevención de la enfermedad.
Siendo ampliamente conocido que los habitantes de medianos y bajos ingresos per cápita a nivel mundial tienen una mortalidad tres veces superior a los de altos ingresos y sólo recibieron el 19,9% de los recursos destinados a diabetes durante el año 2012. El 77% de las personas con diabetes vive en países de ingresos medianos y bajos (Guzmán Rodríguez, 2016).

La diabetes mellitus tipo 2, (DM2) afecta a cerca del 8.3% de la población adulta en todo el mundo, y se estima que ascienda desde 422 millones de personas en el año 2014 a 552 millones de personas para el año 2030 (Whiting, Guariguata, Weil, & Shaw, 2011), lo cual constituye un problema en crecimiento para la salud pública general a nivel mundial.
A pesar de que la diabetes tipo 2 (DM2) es una enfermedad cada vez más frecuente; alrededor de la mitad de los sujetos que la padecen no están diagnosticados (Saaristo, y otros, 2005).

Además, hay 318 millones de personas que padecen de intolerancia a la glucosa que se corresponde con un estadío prediabético que, a su vez, aumenta el riesgo de desarrollar diabetes entre un 5 y 10% en comparación con la población mundial sana que tiene un riesgo promedio de 0.7%. Debido a esto algunos autores en la literatura la consideran la epidemia del siglo XXI a nivel mundial.
Uno de los principales problemas epidemiológicos en este sentido es que apenas la mitad de los casos de diabetes son diagnosticados en forma oportuna, aumentando de esta manera las probabilidades de desarrollo de complicaciones micro y macro vasculares en los pacientes y generando de esta forma una erogación económica para el sistema de salud notable la cual alcanza a representar entre el 5 y el 20% del presupuesto nacional para diabetes (Morsanutto, y otros, 2006).
Dada la magnitud de la lucha contra la DM2, las estrategias implementadas abarcan desde la prevención primaria hasta la detección temprana ambas han demostrado ser efectivas (Guzmán Rodríguez, 2016). Debido a ello, disponer de herramientas diagnósticas simples, efectivas y de bajo costo es de suma importancia. El cuestionario FINDRISC desarrollado y utilizado por los finlandeses para su estudio de prevención primaria de diabetes es un buen ejemplo de ellas (Gagliardino, 2016).

Debido al déficit en términos de diagnóstico recientemente expuesto, se han desarrollado herramientas de cribado tanto para detectar como para predecir el riesgo de padecer esta enfermedad. Actualmente existen varias estrategias para identificar los pacientes con alto riesgo de desarrollar DM2; sin embargo, la de mayor aceptación a nivel mundial es sin duda

la implementación del cuestionario FINDRISC que surgió a partir de un estudio en el cual se consideraron dos cohortes de pacientes una en 1987 y otra en 1992 en Finlandia y se realizó un seguimiento de 10 años de los mismos. El punto de corte sugerido en el estudio original fue de 9 puntos y logró una sensibilidad del 78% y una especificidad del 77% en la primera cohorte de pacientes y de 81% y 76% para la segunda cohorte.

La prueba de FINDRISC (Finnish Diabetes Risk Score) es una herramienta sencilla que, a partir de las respuestas a ocho simples preguntas, permite evaluar el riesgo de una persona a desarrollar DM2 en los próximos 10 años (Lindstrom & Tuomilehto, 2003).

Estas preguntas se refieren a la edad, el índice de masa corporal (IMC), el perímetro de cintura, la actividad física, los hábitos alimentarios, el uso de medicación antihipertensiva, la existencia de registros personales de cifras elevadas de glucemia y a la existencia de antecedentes familiares de diabetes.

El cuestionario FINDRISC reúne varias condiciones que lo hacen atractivo: es simple, no requiere entrevista ya que es autoadministrado, la única intervención auxiliar es la determinación de la circunferencia de cintura, es de muy bajo costo (papel y lápiz), ha sido validado en distintas etnias y condiciones socioculturales y permite alcanzar una primera aproximación a la estimación del riesgo de desarrollar diabetes (Gagliardino J. J., 2016).

El FINDRISC también permite identificar a las personas que tienen un alto riesgo (puntuación entre 15 y 20) o muy alto riesgo (puntuación mayor a 20) de padecer DM2 en los próximos diez años, aún si registran mediciones de valores glucémicos normales.

Esta primera orientación permite seleccionar a los individuos con quienes se debería poner en práctica la intervención en aras de motivar una modificación en su estilo de vida. Esta intervención es la verdadera “Prevención primaria de la DM2”. La efectividad de esta intervención ha sido demostrada ya en diversos estudios previos, con reducciones del riesgo relativo (RRR) que van desde el 39 al 63%, con seguimientos de 3 a 5 años y un número necesario a tratar (NNT) = 4 y 10 pacientes año (Alberti KG, 2007).

El riesgo establecido por FINDRISC para desarrollar DM2 a 10 años de acuerdo con el puntaje del score de la encuesta se agrupa en la Encuesta Nacional de Factores de Riesgo (ENFR) hasta el 2018 en las siguientes categorías:

- Muy bajo riesgo: menos de 7 puntos (probabilidad estimada* de 1/100)
- Bajo riesgo: de 7 a 11 puntos (probabilidad estimada de 1/25)
- Riesgo moderado: de 12 a 14 puntos (probabilidad estimada de 1/6)
- Alto riesgo: de 15 a 20 puntos (probabilidad estimada de 1/3)
- Muy alto riesgo: de 21 a 26 puntos (probabilidad estimada de 1/2)

*De desarrollar DM2 a 10 años.

Si bien la estimación de FINDRISC se basa en un modelo de regresión logística los insumos de este son variables continuas como edad e IMC y categóricas como los antecedentes familiares o la medicación para la hipertensión. En la Encuesta Nacional de Factores de Riesgo (ENFR) se presentan detalles metodológicos de adaptación para cada una de las variables que incluyen discretizaciones de variables continuas y asignación de valores a variables categóricas.

La detección de probabilidades altas de desarrollar DM2 brinda una oportunidad para la detección de la enfermedad o la prevención de complicaciones propias de la misma, así como del manejo temprano de la patología, dado que se trata de una enfermedad susceptible de control.

Cabe destacar que el estado previo al desarrollo de esta patología constituye una etapa fundamental para realizar intervenciones multidisciplinarias que demoren la declaración de la enfermedad y reduzcan costos humanos y económicos derivados necesariamente de las complicaciones propias de la patología.

La variabilidad de las diferentes regiones debido a componentes culturales, así como factores genéticos han motivado que en el transcurso de las últimas dos décadas se realizarán varios estudios con la finalidad de validar dicha encuesta en poblaciones específicas y en varias de ellas se han propuesto distintos puntos de corte del cuestionario, así como de medidas como el perímetro de cintura.

La ENFR implementada en nuestro país durante el periodo 2005-2013, reportan que la prevalencia de diabetes en nuestra población adulta aumentó del 8,4 al 9,8%, por lo que existirían en Argentina más de 2 millones de personas con diabetes, representada por su forma clínica más común: la diabetes tipo 2 (DMT2) (Ministerio de Salud de la Nación, Instituto Nacional de Esadísticas y Censos, 2015).

En nuestro país, la prevalencia de diabetes o glucemia elevada por auto reporte aumentó desde 9.8% en 2013 a 12.7% en 2018 y según la (ENFR) la prevalencia combinada de ambas en 2018 resultó de 10.9%.

Para comprender mejor el panorama sanitario de nuestro país, falta hay que destacar que entre los encuestados que no se auto reportaron como diabéticos ni refirieron haber tenido un registro previo de glucemia elevada, el 5% obtuvo mediciones objetivas de glucemia elevada y solamente la mitad de los diabéticos ya diagnosticados estaban recibiendo algún tipo de tratamiento médico.

Por último, al estimar el riesgo de diabetes se encontró en nuestra población que casi un 20% presenta un alto riesgo de desarrollar DM2 a 10 años.

1.2. Objetivos del Trabajo

Principal

El objetivo principal de este estudio es desarrollar dos modelos para predecir el nivel de Glucemia en ayunas. El primero utilizando los pacientes que han respondido la encuesta Findrisc y el segundo a partir de una base de laboratorio de análisis clínicos.
Si bien en principio se busca estimar la incidencia de diabetes y de riesgo de desarrollarla a diez años en una población de la región argentina de Cuyo, también es importante este dato como factor de riesgo cardiovascular en esta población.

Secundarios

- Comparar el desempeño de diferentes modelos de regresión clásicos y propios de aprendizaje automático utilizando diferentes métricas.
- Ajustar los hiperparámetros de los modelos para lograre el mejor desempeño posible respecto de estas métricas.
- Construir ensambles de modelos a fin de optimizar la precisión de las predicciones.

2. Capítulo 2: El Caso de estudio

2.1. Génesis de la base de datos de Findrisc

Se ha construido una base de datos con los registros correspondientes a los afiliados de la obra social OSEP que cubre a los empleados públicos de la provincia de Mendoza, Argentina, algunos de los cuales respondieron el cuestionario de FINDRISC. La muestra fue recolectada durante el período que se inicia en enero de 2018 hasta diciembre de 2022, y en ella se incluyen indicadores de riesgo a 10 años y también la estimación de la edad vascular.

Algunos de los datos de la base corresponden a variables que fueron obtenidas de la encuesta de FINDRISC (ver anexo 1), tales como edad, índice de masa corporal (IMC, tabaquismo y, antecedentes de diabetes, entre otras, así como variables que se obtuvieron en el momento de la recolección de la muestra, como la presión arterial (tanto diastólica como sistólica), peso, talla, perímetro de cintura, ocupación, fecha de nacimiento y lugar de residencia. Además, la base de datos incluye indicadores de edad vascular y la estimación del riesgo a 10 años de padecer DM2.

En algunos casos se cuenta con variables adicionales, tales como niveles de glucemia (292 registros) y colesterol total (207 registros).

2.2. Génesis de la base de datos de Laboratorio

Este archivo se solicitó en forma adicional a OSEP con el propósito de estimar la glucemia en ayunas de los pacientes en los cuales falta este dato y resulta fundamental para el objetivo del presente trabajo. Esta base está formada por 628 registros correspondientes a 63 variables y contiene información adicional de análisis de laboratorio desde Enero de 2021. Existe un grupo de personas que figuran al mismo tiempo en la base de datos de Laboratorio y en la base de Findrisc.

Es importante destacar que algunos de los registros corresponden a pacientes ambulatorios mientras que otros son de pacientes internados. En el caso de pacientes internados puede haber más de un registro en un mismo día.

Respecto de esta base complementaria cabe mencionar que:

- De las observaciones de glucemia 401 corresponden a pacientes ambulatorios, mientras que 22 corresponden a pacientes internados.

- Se tienen 131 mediciones de Hemoglobina Glicosilada (Hba1c) que es un indicador importante para síndrome metabólico y/o diabetes.
- Se tienen 320 determinaciones de Colesterol total y de triglicéridos
- Se tienen 315 mediciones de HDL que facilitarían el diagnóstico del Síndrome Metabólico.

En la base de Laboratorio se incluyen variables sociodemográficas como la edad y el sexo además de mediciones de laboratorio como niveles de hemoglobina, glucemia, creatinina en sangre, sodio, potasio, bilirrubina, colesterol, entre otras. Dentro de las variables hay algunas de tipo categórico y otras de tipo cuantitativo.

También debe destacarse que hay una gran cantidad de variables con datos faltantes, lo que puede dificultar el análisis de algunos aspectos de los datos. Esta particularidad es fundamental para considerarla ya que puede que algunas metodologías no funciones bien en presencia de gran cantidad de valores faltantes.

Se excluyeron del análisis las variables no vinculadas directamente con el objetivo de este trabajo y se concentró la atención en las características más representativas con pocos valores nulos para utilizarlas en la ingeniería de atributos.

2.3. Análisis Exploratorio de Datos

2.3.1. Findrisc

La base de datos principal de estudio contiene información de 2163 asociados a OSEP, quienes completaron el formulario FINDRISC que consta de 26 variables, 11 de ellas categóricas y 15 numéricas. La mayoría de las variables presentan valores faltantes, excepto Findrisc, Edad vascular, estimación del riesgo a 10 años y fecha, que están completas.

Las variables con mayor cantidad de registros nulos son Observaciones (2121 casos), Colesterol (1986 registros) y Glucemia (1871 registros). Ocupación cuenta con 353 valores faltantes, mientras que TA diastólica cuenta con 132. Las demás variables tienen menos de 30 registros nulos, como Sexo (20), fecha de nacimiento (22), Edad (23), Peso (20), Talla (20), IMC (20), perímetro de cintura (28), TA Sist. (28), Actividad Física (23), Vegetales/Frutas (22), Hipertensión (25), Trat. HTA (24), Glucosa (Si/No) (28), Antecedentes de Diabetes (29), Diabetes (24) y Fumador (26).

Con el objetivo de maximizar la información disponible para el entrenamiento de los algoritmos, se convierten las variables categóricas Sexo, Actividad física, Hipertensión, Tratamiento HTA, Glucosa (Si/no), Antecedentes de diabetes, Diabetes y Fumador en variables ficticias (dummies). Se generan las variables Sexo_F y Sexo_M a partir de Sexo, del mismo modo se procede con Activ.Fisica, Hábito de tabaquismo, Vegetales/Frutas todos los días, Hipertensión se crean las variables y Tratamiento HTA. Para Antecedentes de Diabetes se consideran tres niveles: Antec.Diabetes_NO, Antec.Diabetes_Sí (abuelos, tíos o primos hermanos) y Antec.Diabetes Sí (padres, hermanos o hijos propios).

2.3.1.1. Asociación entre variables

Para identificar las variables con mayores niveles de asociación en Findrisc utilizaremos una medida estadística no paramétrica llamada coeficiente de correlación de Spearman. Dado que la base de datos de Findrisc contiene valores atípicos, anómalos y extremos es conveniente elegir una medida robusta que no se vea afectada por estos registros y puesto que la normalidad bivariada tampoco se satisface no es posible aplicar el test de correlación de Pearson y es más adecuada la aplicación de esta prueba no paramétrica.

En la Figura 1 se presenta una selección de las variables más correlacionadas con Findrisc de la base y mediante un correlograma se visualiza la fuerza de esta asociación de un modo más compacto:

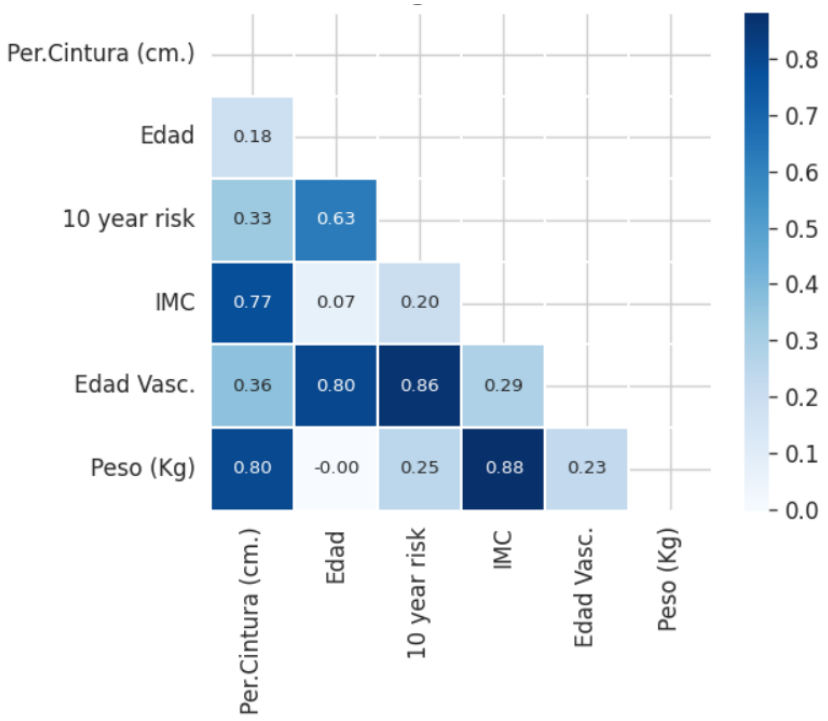


Figura 1: Asociación entre variables continuas de la base de Findrisc

Las variables de peso e IMC (Índice de Masa Corporal) es un indicador de obesidad, las personas tienen un peso normal si su IMC se encuentra entre 18 y 25, un IMC superior a 25 pero inferior a 30 se considera como sobrepeso y los superiores a 30 se consideran como obesos. Dentro de la obesidad a su vez se distinguen varios grados. Existen otros indicadores similares como el perímetro de cintura o superficie corporal que se utilizan para estimar riesgo cardiovascular. El IMC se define como el cociente entre el peso en kg y el cuadrado de la estatura en metros.

Las variables Edad Vascular y Riesgo a 10 años de enfermedad cardiovascular están fuertemente correlacionadas dado que ambas se vinculan con el envejecimiento y el deterioro de la salud cardiovascular. Esto es lógico dado que sus expresiones se basan en las mismas variables; Edad, Diabetes, Tabaquismo, Hipertensión, IMC.

2.3.1.2. Calidad de datos

Se verificaron los valores calculados de las variables de riesgo a 10 años y edad vascular en el conjunto de datos. Estas variables son importantes para predecir el riesgo de enfermedad cardiovascular a largo plazo y sugerir medidas preventivas. Las ecuaciones utilizadas para calcular estas variables se basan en factores de riesgo cardiovascular, como la edad, el género, el colesterol, la presión arterial y el tabaquismo.

Creamos dos funciones, una llamada Edad Vascular y la otra llamada Riesgo a 10 Años. Como primera medida tenemos que pasar las variables binarias Diabetes (D) y Fumador (F). Para estimar riesgo a 10 años, se utilizaron los algoritmos propuestos por la Asociación Americana de Riesgo Cardiovascular en su proyecto Lifetime Risk Pooling (Bundy, 2020).

Se estudió la diferencia entre el valor original disponible en la base y el valor estimado mediante las expresiones presentadas. Los resultados del estudio indicaron que las diferencias entre los valores originales y los recalculados no eran significativas, con una distribución entre -0.004 y 0.004 años. Por lo tanto, se llegó a la conclusión de que los datos en ambas variables pueden considerarse correctos. La distribución de las diferencias entre los valores de la edad vascular reportados y calculados en esta población se puede apreciar en la Figura 2.

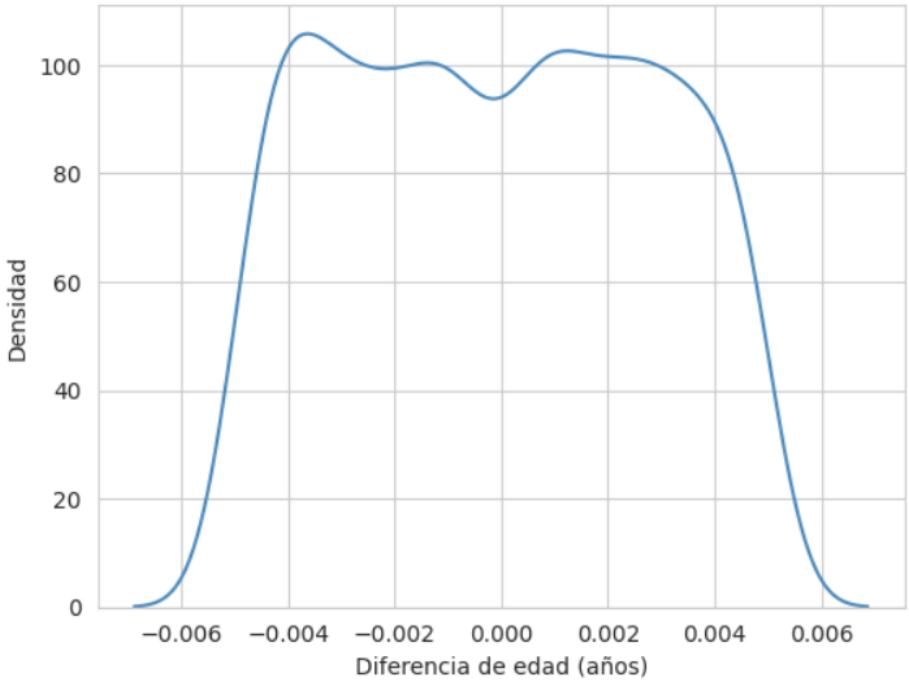


Figura 2: Distribución de las diferencias entre la variable edad vascular calculada y reportada

Para Riesgo a 10 años se visualizaría de la siguiente forma:

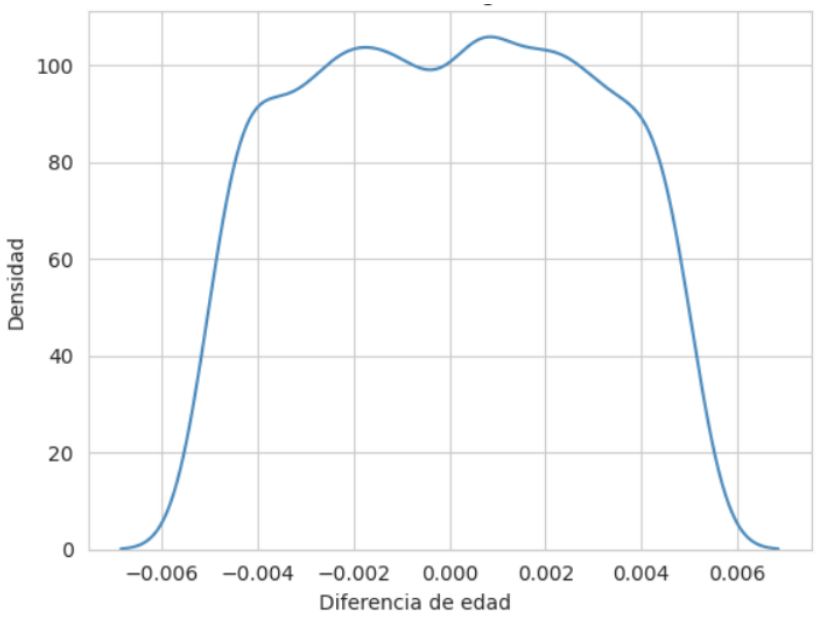


Figura 3: Distribución de las diferencias entre la variable riesgo a 10 años calculada y reportada

Para garantizar la consistencia de la información, así como la integridad y confiabilidad del análisis y la posibilidad de generalizar los resultados se intentaron rescatar los registros cargados con error en la variable glucemia utilizando técnicas de corrección, validación y normalización.

2.3.1.3. Análisis de Bland Altman

Es usual en el campo de la salud buscar metodologías alternativas de cuantificación, ya sea para mejorar la precisión de un método o bien para contar con una alternativa que de alguna forma consuma menor cantidad de recursos. Un ejemplo clásico que podría citarse es cuando se comenzó a utilizar la evaluación de presión arterial con esfigmomanómetros digitales, y se comparaban sus resultados con un esfigmomanómetro tradicional de mercurio. Lo que se pretende responder es si los dos métodos son equivalentes, es decir si concuerdan lo suficientemente bien o no en sus mediciones (Cardemil, 2017).

En 1983, Altman y Bland plantearon su visión respecto a la comparación de dos métodos, fundamentando esta propuesta en que el análisis con correlación de Pearson no es del todo adecuado, ya que no detecta una diferencia sistemática de nivel entre ambas mediciones. Para ello mostraron que la correlación de Pearson no evalúa concordancia entre las mediciones, sino que evalúa asociación lineal entre las mediciones (variables), por lo que

dos métodos pueden correlacionarse muy bien, pero sin embargo concordar muy poco (Cardemil, 2017).

Altman y Bland proponen un análisis gráfico que permite comparar dos técnicas de medición sobre una misma variable cuantitativa o bien evaluar la reproducibilidad de un solo método. Es muy apropiado cuando se quiere validar un nuevo método de medición. Este método cuantifica la diferencia media entre ambos métodos y estima un rango de confianza de este sesgo que se espera incluya al 95% de las diferencias entre ambas mediciones (Giavarina, 2015). Idealmente el sesgo debería ser nulo.

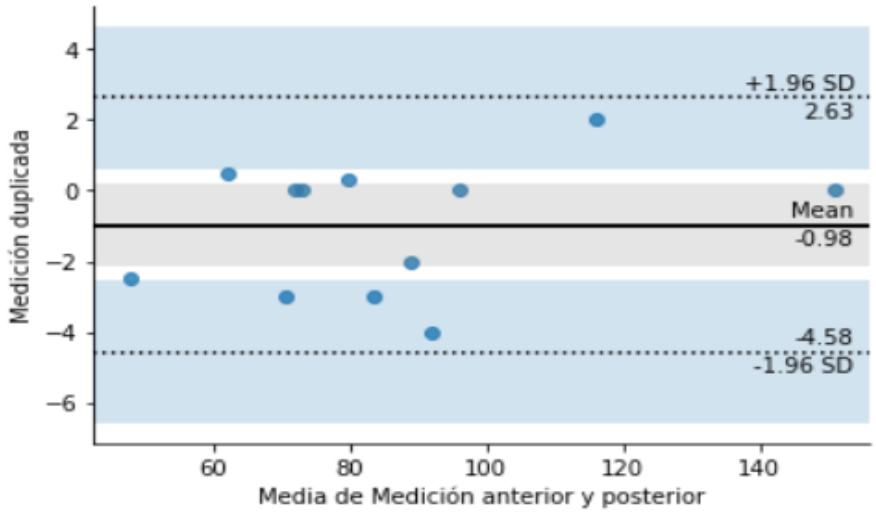


Figura 4: Gráfico de Bland Altman para la variable Peso

Las tres líneas horizontales graficadas representan:

Límite de concordancia superior: es el límite superior del intervalo de confianza del 95% de la diferencia estimada entre ambos métodos de medición, suponiendo normalidad asintótica; es decir apoyándose en el Teorema Central del Límite.

Diferencia media: es el promedio de las diferencias observadas entre ambas metodologías y da cuenta del error sistemático, en caso de ser no nula.

Límite de concordancia inferior: es el límite inferior del intervalo de confianza del 95% de la diferencia estimada entre ambos métodos de medición.

Cuanto menor resulte el rango entre los límites de concordancia de la gráfica, mayor será el grado de acuerdo de ambas mediciones.

En el caso de que ambos métodos produzcan mediciones similares, las diferencias se ubicarán en torno a una línea horizontal de nivel 0.

Si los valores de las diferencias se alejan mucho de esta línea se tendrá una señal de que ambos métodos producen mediciones diferentes. Esto puede indicar que el nuevo método sobreestima o subestima el valor objetivo.
Asimismo, es importante observar si la variabilidad es constante a lo largo de todo el rango de valores de las mediciones o bien si esta variabilidad es mayor en alguna parte del recorrido de la variable.
En el análisis para el conjunto de datos de Laboratorio se detectaron 9 pares de valores duplicados por ID, es decir, con distintas mediciones para cada una de las variables para una misma persona. Para analizar si los valores registrados para cada una de las variables son concordantes bajo el supuesto que se realizaron bajo diferentes técnicas de medición, se decidió realizar un modelo de Bland-Altman para cada una de las variables numéricas de la base de datos que registren diferentes valores para una misma persona.

Para realizar el análisis se construyó una base de datos auxiliar con los valores duplicados por la variable “ID” y por la variable “Fecha” y se ordenaron los registros de menor a mayor, utilizando como definición que las diferentes técnicas de medición se utilizaron en diferentes oportunidades. Luego se subdividió esta base en dos subbases una con los registros pares y otra con los registros impares, a fin de concatenar ambas por variable y evaluar las diferencias.
Del análisis exhaustivo de cada una de las variables no surge rechazo de la hipótesis de nulidad que establece que ambos métodos de medición producen resultados similares a causa de que, en la gran mayoría de los casos, la media se encontraba próxima a 0, los puntos se situaban dentro de los intervalos de confianza de las diferencias y las distribuciones presentaban un grado de dispersión uniforme.

A continuación, se seleccionaron los mejores registros para cada ID, considerando previamente si alguno de ellos contenía mayor cantidad de información, es decir si uno de los registros presentaba datos en una mayor cantidad de variables que en el otro, pero no, si bien los registros difieren en sus valores los pares de duplicados vienen completos en las mismas variables. Por lo tanto, al ser descartada la posibilidad de acrecentar el nivel de información y al estar seguro de que ambos métodos producen información concordante se optó por tomar una definición en cuanto a la fecha de realización del estudio tomando el más actual como el registro que prevalecerá sobre el otro para una misma persona.

2.3.1.4. Detección de Anomalías y Outliers

Se conoce como outliers o valores extremos a aquellos datos que están muy alejados del conjunto general. Estos datos pueden ser muy relevantes en el análisis y existe una variedad de metodologías denominadas robustas para integrar estos datos al análisis reduciendo al mínimo el sesgo que estos producen en la estimación y la inferencia.

Los outliers pueden deberse a errores de carga o bien a atipicidades características de las propias observaciones. Si se trata de un error de carga, puede buscarse el valor de ese dato o bien imputarse por cercanía con otros o bien trabajar con ese dato faltante. El modo de imputación más conveniente dependerá de la metodología que se pretenda aplicar. Estas atipicidades pueden detectarse en una variable en particular y en ese caso la detección será sencilla, o bien pueden deberse a la forma de asociación de un conjunto de variable y en forma univariada no es posible encontrarlas.

Asimismo, pueden presentarse en forma aislada o bien en agrupaciones. Cuando se observa un patrón diferente y varios datos agrupados con este formato estamos en presencia de una anomalía. En la literatura se han propuesto diversas metodologías para detectar tanto outliers como anomalías. Algunas son adecuadas para datos estructurados y también las hay para datos no estructurados.

Isolation Forest

Para la detección de valores atípicos una de las metodologías que se ha empleado es el método llamado Isolation Forest. Este método se basa en la idea de que los valores atípicos son más fáciles de aislar que los datos normales. Isolation Forest utiliza árboles aleatorios para dividir el conjunto de datos en subconjuntos más pequeños y, por lo tanto, identificar los valores atípicos que quedan aislados en las ramas más cortas del árbol. La ventaja del Isolation Forest es su capacidad para manejar grandes conjuntos de datos y detectar outliers de manera eficiente. También es adecuado para datos con alta dimensión, lo que lo hace particularmente útil para muchos casos de uso en la ciencia de datos. Entre sus desventajas, se encuentra que no es muy preciso cuando la tasa de outliers es alta en comparación con la tasa de datos normales. Además, si los outliers están muy cerca de los datos normales, no es recomendable la utilización de este método.

Bajo esta modalidad, si bien se detecta la presencia de valores atípicos en las respectivas variables, es recomendable a posteriori realizar un análisis pormenorizado. En nuestro caso concluimos que ninguno de ellos corresponde a errores de carga, es decir, son valores aceptables para esa variable, por lo cual no corresponde eliminar o reemplazar valores. A

modo de presentación de resultados del citado método podemos graficar las siguientes variables:

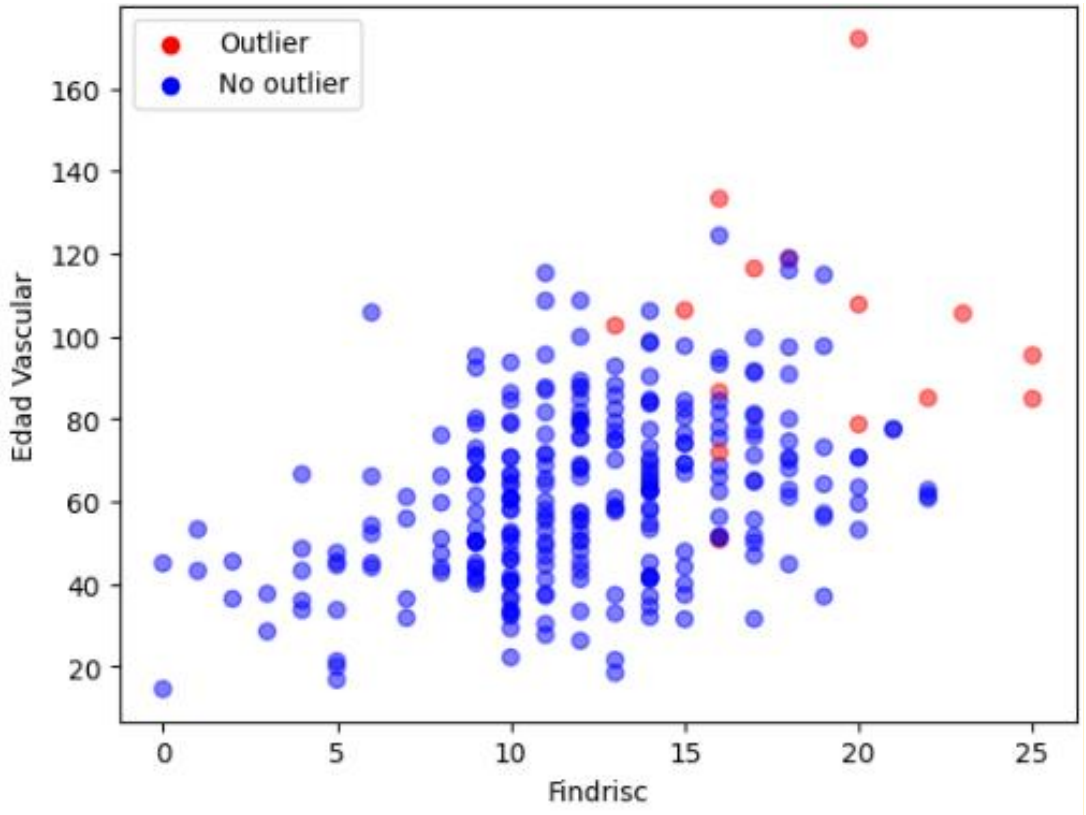


Figura 5: Detección de anomalías por Isolation proyección sobre el diagrama de dispersión de Edad Vascular y Findrisc

Se aprecia en la Figura 5 que la posición de los valores anómalos se ubica en pacientes con alto riesgo señalado por el test de Findrisc y además con edad vascular avanzada.

Local Outlier Factor

Para la detección multivariada de outliers utilizaremos el método no supervisado de Local Outlier Factor (LOF). El LOF es un método de detección de outliers que se basa en la densidad de los datos. El procedimiento de cálculo consiste en seleccionar un punto de los datos y realizar el cálculo de su vecindad inmediata, considerando un radio de búsqueda de alrededor de ese punto, lo que implica encontrar el número de puntos que caen dentro de ese radio. La idea detrás de LOF es que los outliers son aquellos puntos que tienen una densidad significativamente menor en comparación con sus vecinos cercanos. Un objeto con un LOF alto se considera un outlier.

Entre sus ventajas podemos mencionar que no depende de la distribución de los datos, puede manejar conjuntos de alta dimensionalidad y puede captar relaciones no lineales entre datos. Entre las desventajas se encuentra su sensibilidad a la elección de sus parámetros, como el número de vecinos cercanos, que pueden afectar significativamente los resultados, puede tener dificultades para manejar datos con distribuciones muy heterogéneas o densidades muy diferentes y por último puede ser costoso computacionalmente en grandes conjuntos de datos debido a la necesidad de calcular distancias y vecindarios. En la Figura 8 se representa la detección de outliers por LOF proyectados sobre el diagrama de dispersión de Glucemia versus Creatinina en sangre.

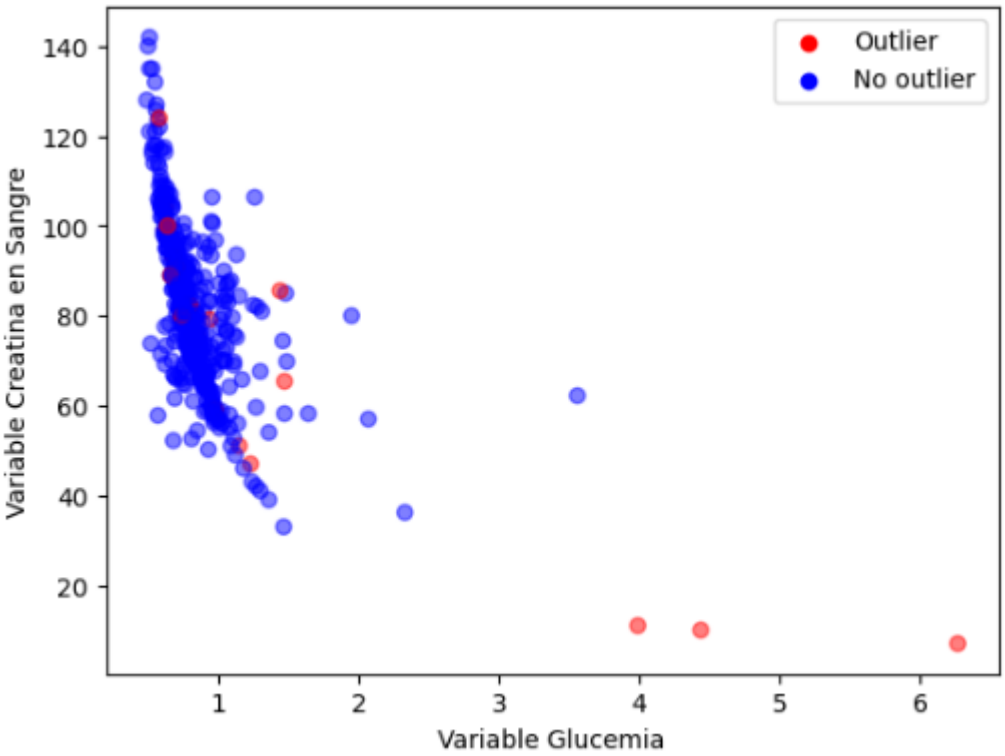


Figura 6: Local Outlier Factor entre la variable Creatina en Sangre y Glucemia

Se aprecia que algunos puntos corresponden a valores altos de glucemia que probablemente indican condición diabética, otros corresponden a valores altos de creatinina que indica probablemente deterioro de la función renal.

Criterio de Tukey

También, siguiendo el criterio de Tukey para la detección de valores atípicos univariados, se hallaron valores fuera de rango en las variables Transaminasa glutámico-oxalacética (TGO o AST) y transaminasa glutámico-pirúvica (GPT o ALT), con 23 casos y hemoglobina

en(g/dL) por encima de 20 casos. El resto de las variables tienen un número relativamente bajo de valores atípicos, siendo 7 variables con valores mayores a 10 casos y 12 menores a 10. las dos variables que no cuentan con anomalías son edad y MDRD estimado para hombres.

De acuerdo con el criterio médico, estos datos atípicos deben considerarse válidos, asumiendo que en el momento de realizar la extracción el paciente estaba cursando alguna patología que generará esos valores, con el objetivo de representar esas situaciones y que el modelo las considere posibles, no se realizará ningún tratamiento a los mismos.

2.3.2. Laboratorio

La base de datos auxiliar en este estudio correspondiente a "Laboratorio" contiene 628 registros con 25 variables de distintas mediciones que se utilizan con el objetivo de diagnosticar Síndrome metabólico. El síndrome metabólico es el nombre que se da a un grupo de factores de riesgo de enfermedad cardiaca o diabetes y otros problemas de salud. Un paciente puede tener un solo factor de riesgo, pero, a menudo, los individuos presentan varios simultáneamente. Estos factores incluyen: obesidad abdominal, es decir demasiada grasa acumulada alrededor de la cintura, nivel alto de triglicéridos, nivel bajo de colesterol HDL, alta presión arterial y nivel alto de glucemia en ayunas. Cuantos más factores de riesgo se tengan mayor es el riesgo de enfermedad cardiovascular, diabetes o accidente cerebrovascular.

Tenemos un ID por paciente coincidente con algunos registros de la base de Findrisc, un campo temporal de la fecha de estudio, 5 campos categóricos y los ya mencionados 25 campos numéricos.
Todas las variables excepto ID, Fecha, Sexo, Extracción, Procesamiento, Situación, Edad presentan un alto porcentaje de valores nulos destacándose entre ellas la variable MDRD_hombre_na (se refiere a la tasa de filtración glomerular estimada (TFGe) para un hombre utilizando la ecuación MDRD (Modification Diet in Renal Disease) con 576 registros nulos. La presencia de esta cantidad de datos nulos en la base representa un desafío importante en el proceso de modelado debido a que dificulta la obtención de métricas precisas, lo que finalmente se traduce en una limitación en la capacidad para extraer información valiosa de los datos, dentro del apartado de modelado se da tratamiento a esta problemática.
Dentro de las variables categóricas la única que aporta poder predictivo para la construcción del modelo es la variable Sexo, por lo tanto, las restantes se descartan para el análisis.

2.3.2.1. Asociación entre variables

Utilizaremos el coeficiente de correlación de Spearman como medida de asociación entre variables dado que no cumplen los supuestos de normalidad bivariada, por lo cual no es adecuado la prueba paramétrica de Pearson (Zhang, 2018).
Seleccionamos las variables con correlaciones más significativas con el fin de poder visualizarlas de una forma más compacta en un correlograma:

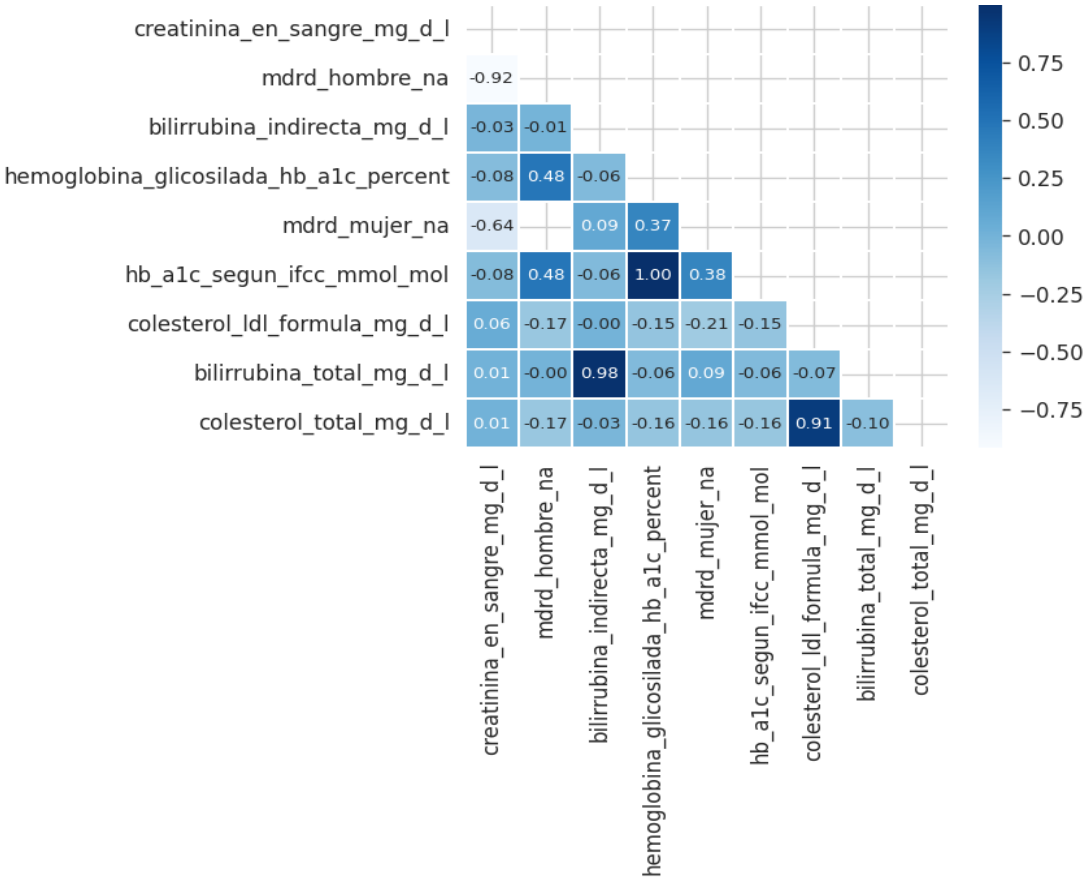


Figura 7: Correlaciones significativas para el data set de Laboratorio

Si bien son notables algunas asociaciones, esto se debe a una correlación biológica entre las mediciones o bien a que son mediciones en diferente unidad de la misma variable. Tal es el caso de la hemoglobina glicosilada HbA1c que proporciona información sobre los niveles de glucosa en sangre promedio durante los últimos tres meses. Esta variable puede medirse utilizando distintas unidades de medición y es esperable que estas cuantificaciones se hallen altamente correlacionadas. Una observación análoga puede hacerse respecto a las mediciones de bilirrubina indirecta y bilirrubina total dada su alta correlación fisiológica.

Respecto a la fuerte correlación observada entre la tasa de filtración glomerular estimada por MDRD y los niveles de creatinina en sangre puede deberse a que la creatinina en sangre

se utiliza para la estimación del MDRD y otros indicadores de filtrado glomerular por lo tanto es razonable que estén fuertemente asociados estos valores.
Finalmente, la alta correlación entre las concentraciones de colesterol total y colesterol LDL (lipoproteína de baja densidad) en sangre se deben a la definición del colesterol total.

Cabe destacar que ambas mediciones son indicadores de la salud cardiovascular y del metabolismo de lípidos en el cuerpo. Es importante destacar que, aunque la correlación entre estas dos variables es alta, es necesario evaluar otras variables y factores de riesgo para una evaluación completa del perfil lipídico y el riesgo cardiovascular.

2.3.2.2. Calidad de datos

Para que la inferencia basada en el modelo de regresión lineal resulte válida es necesario que se cumplan los supuestos del modelo lineal. Estos supuestos se refieren a la distribución de los residuos del modelo e incluyen la normalidad distribucional, la media nula, la homocedasticidad y la independencia de las observaciones. Por propiedades estadísticas, se desprende del cumplimiento de estos supuestos que la variable respuesta también tiene distribución normal.

Dada la importancia del registro de glucemia en ayunas para el diagnóstico de diabetes, así como para la estimación del riesgo de desarrollarla y siendo este un dato con muchos faltantes en la base original se pretende a partir de la base complementaria de datos de laboratorio estimar la glucemia de los pacientes de la base Findrisc. Sin embargo, no es posible sostener el supuesto distribucional de normalidad para esta medida en esta población, esto puede apreciarse en la Figura 7 donde se evidencia una marcada asimetría por la derecha en la distribución de esta variable.

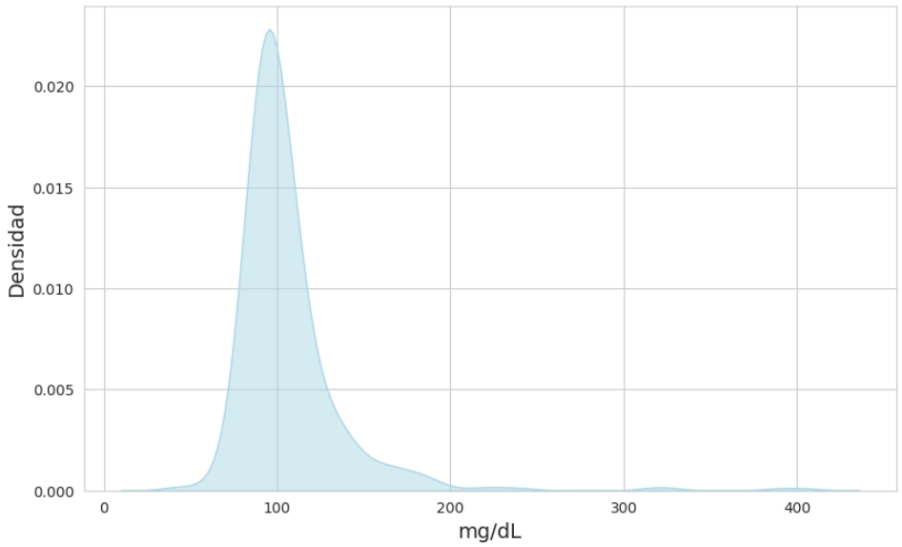


Figura 8: Distribución de la variable Glucemia

Si bien no es necesario corregir datos de la base de laboratorio dado que todos se mantienen dentro de los límites razonables para las variables, será necesario atender a la particularidad de ser una base con un alto porcentaje de datos faltantes al momento de considerar las metodologías adecuadas para este contexto.

3. Capítulo 3: Los Modelos

3.1. Evaluación del desempeño

Para comparar el desempeño logrado por los distintos modelos de regresión alternativos nos planteamos considerar distintas medidas de bondad del ajuste para los modelos. Entre ellas vamos a considerar el R^2 ajustado y otras métricas como el RMSE, MAE y MAPE. La expresión para el coeficiente R^2 ajustado se define como:

$$\overline{R^2} = 1 - [(1 - R^2) * \frac{(n-1)}{(n-k-1)}]$$

Dónde:
R^2 es el coeficiente de determinación simple
n es el número de observaciones en la muestra
k es el número de variables independientes (predictoras) en el modelo

El coeficiente R^2 ajustado se utiliza para evaluar el grado de ajuste de un modelo de regresión, y penaliza la inclusión de variables que no aportan información significativa al modelo, si bien R^2 aumenta con la cantidad de variables, R^2 ajustado no necesariamente sigue ese comportamiento dada la penalización de la cantidad de variables incluidas en el modelo. Como consecuencia, y según (Smith, 2018), el coeficiente de determinación R^2 ajustado es una medida más confiable que el coeficiente de determinación R^2 en modelos con múltiples variables predictoras. Cuanto mayor sea el valor de R^2 ajustado, mejor será el ajuste del modelo.

La raíz cuadrada del error cuadrático medio (RMSE, por sus siglas en inglés), se calcula como la raíz cuadrada del Error Cuadrático Medio (Mean Squared Error, MSE), es decir, la raíz cuadrada de la media de los cuadrados de las diferencias entre las predicciones del modelo y los correspondientes valores observados. Se define como:

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

Indica el ajuste absoluto del modelo a los datos, cuán cerca o cuán lejos están los puntos observados de los valores predichos del modelo. RMSE se puede interpretar como la desviación estándar de la varianza no explicada por el modelo (Hastie, 2009). Esta medida está siempre expresada en las unidades de la variable respuesta y es no negativa, pero su interpretación no es tan intuitiva como la del MAE y se ha mostrado que presenta cierta sensibilidad a los valores atípicos en los datos.
Además, penaliza severamente los errores de predicción, es decir las predicciones que se alejan de los datos reales.
En resumen, cuanto menor sea el RMSE, mayor será la precisión del modelo. RMSE es una excelente opción para evaluar la precisión con que el modelo predice la respuesta, y es el criterio más importante para ajustar si el propósito principal del modelo es la predicción.

La siguiente métrica propuesta es el error absoluto medio (Mean Absolute Error, MAE). Se calcula como la media de las diferencias absolutas entre las predicciones del modelo y los valores reales correspondientes, por lo tanto, cuanto menor sea el MAE, mejor será la precisión del modelo (James, 2013). La expresión para el cálculo del MAE es:

$$MAE = \frac{SAE}{N} = \frac{\sum_{t=1}^{N} |y_t - \hat{y}_t|}{N}$$

Esta medida se destaca por la sencillez de su cálculo ya que todos los errores se ponderan en la misma escala, es útil si los datos de entrenamiento tienen valores atípicos, ya que el MAE no penaliza tan fuertemente los errores elevados causados por valores atípicos (por lo tanto compensamos entre ambas métricas), sin embargo no puede utilizarse para comparar series utilizando diferentes medidas y no es diferenciable en cero (lo que puede suponer una dificultad en cuánto a su uso en determinados algoritmos de optimización, aunque puede suponer una ventaja en ciertos casos particulares).

Finalmente, la última métrica a utilizar para evaluar los diferentes modelos es el MAPE (Mean Absolute Percentage Error o Error Porcentual Absoluto Medio en Español) es una medida de error relativa que utiliza valores absolutos para evitar que los errores positivos y negativos se cancelen entre sí (Hyndman, 2006). Su expresión de cálculo es:

$$MAPE = \frac{1}{N}\sum_{t=1}^{N}\frac{|y_t - \hat{y}_t|}{y_t}$$

El MAPE promedia los errores porcentuales absolutos, es decir, el porcentaje del valor real que representa la diferencia entre el valor estimado y el real.

El MAPE es intuitivo y de fácil interpretación, al ser una medida relativa no se ve afectada por el tamaño de los valores absolutos de los datos, convirtiéndola en una métrica particularmente útil para comparar el rendimiento de modelos que trabajan con datos de diferentes escalas y que penaliza de manera equitativa tanto las sobreestimaciones como las subestimaciones. Sin embargo, puede dar resultados incorrectos si hay valores reales nulos. En estos casos convendría utilizar otra métrica como SMAPE.
La expresión del SMAPE es la siguiente:

$$SMAPE = \frac{1}{N}\sum_{i=1}^{n} 2 * \frac{|Y_i - \widehat{Y}_i|}{|Y_i| + |\widehat{Y}_i|}$$

El SMAPE se expresa como porcentaje y tiene la ventaja de ser simétrico, lo que significa que penaliza igualmente tanto las sobreestimaciones como las subestimaciones (Armstrong, 2006). Sin embargo, puede resultar más compleja su interpretación debido a que su escala no es lineal, razón por la cual no la utilizaremos para la evaluación de modelos.

3.2. Modelo de Predicción de Glucemia sobre Laboratorio

Se planea construir un modelo de regresión de cuadrados mínimos ordinarios (Ordinary Least Squares, OLS) con el objeto de predecir el valor de glucemia ausente en gran cantidad de registros a partir de otros valores de laboratorio disponibles.

Para realizar este análisis, se desconsideran todos los valores nulos de la variable objetivo, logrando consolidar una base de 433 registros completos. Si bien hay pacientes que han sido evaluados en más de una oportunidad, en ocasiones por ser pacientes internados, en esta instancia se consideran todas las mediciones a fin de disponer de una mayor cantidad de datos.

En esta oportunidad, se identifican atipicidades en la variable objetivo y se decide realizar una cuidadosa inspección de éstas a fin de determinar si son outliers o bien errores de tipeo y se determina que corresponden a mediciones legítimas pero extremas.

Esto nos conduce a la inclusión de estos valores en la construcción y validación del modelo, con el debido cuidado de que no afecten en demasía la robustez y estabilidad del mismo.

Con el fin de asegurar que todos los modelos puedan ser evaluados en forma objetiva y confiable, se dividirá la totalidad de datos de forma aleatoria, es decir, no estratificada, en un conjunto de entrenamiento (train) y un conjunto de validación (test), dado que no disponemos de grandes cantidades de datos, los porcentajes destinados a cada conjunto serán de 70 % para entrenamiento del modelo de machine learning y 30 % que utilizaremos para la validación.

La división en conjuntos de entrenamiento y prueba es una metodología usual en estadística y en aprendizaje automático pues permite evaluar el grado de ajuste evitando el sobreajuste (overfitting), es decir que el modelo aprenda del ruido de los datos de entrenamiento y pierda capacidad para la generalización (Witten, 2016).

El overfitting se produce cuando un modelo de aprendizaje automático se ajusta demasiado a los datos de entrenamiento y por lo tanto no generaliza bien a nuevos datos no vistos. Para identificarlo, se divide a los datos en un conjunto de prueba y otro de testeo, si el modelo tiene un rendimiento más desfavorable en el conjunto de validación que en el conjunto de entrenamiento es altamente probable que se haya producido overfitting. Entre las técnicas para evitarlo se encuentran: la regularización (consiste en agregar una penalización a la complejidad del modelo durante el entrenamiento), el uso de una mayor cantidad de datos o técnicas de validación cruzada (implica dividir el conjunto de datos en múltiples partes o folds y realizar múltiples iteraciones de entrenamiento y validación (Goodfellow, 2016).

Se desconsideran, en esta parte del análisis, las variables de fecha de procesamiento, extracción, identificación del paciente y situación para la construcción del modelo de regresión dado que no aportan valor predictivo.

Lamentablemente el modelo de regresión de mínimos ordinarios no satisface el supuesto de homocedasticidad (los errores no tienen una varianza constante); además, existe una marcada multicolinealidad entre algunas de las variables predictoras, lo que dificultaría la interpretación de los coeficientes y la fiabilidad de los resultados y, por último, los errores no siguen una distribución normal con media cero. Lo que nos indicaría que el modelo de

regresión de mínimos ordinarios no es adecuado para los datos y es necesario explorar otras alternativas.

Bajo este escenario ensayamos algunos modelos alternativos como modelos de regularización y modelos basados en componentes principales, pero todos resultaron poco adecuados debido a una marcada presencia de valores faltantes, por lo que se procedió a explorar algunos modelos de machine learning. En este sentido, la selección del modelo de machine learning adecuado es una tarea fundamental.
Para una primera evaluación se optó por el modelo Histogram-Based Gradient Boosting. Se trata de un estimador que utiliza un ensamble de árboles de decisión que se basa en la metodología de Gradient Boosting cuyo objetivo final es construir un modelo de regresión preciso y eficiente mediante la combinación de muchos modelos de regresión más simples, denominados "modelos débiles" (Friedman, 2001).

La metodología es iterativa y ajusta un nuevo modelo a los residuos del anterior en forma secuencial, permitiendo mejorar la precisión del ajuste del modelo global. En este caso utiliza una técnica de histograma para agrupar y procesar los datos de manera eficiente, lo que permite que el modelo pueda manejar grandes conjuntos de datos y características continuas sin afectar el rendimiento.

El resultado del modelo nos arroja el siguiente cuadro de resultados:

Tabla 1: Resumen métricas Histogram-Based Gradient Boosting

Modelos	R2 ajustado	MAPE	MAE	RMSE
HistGradientBoosting sin imputación	0,27	13,76%	15,35	23,81

Siendo estos valores característicos de un ajuste razonable pero optimizable. Si bien Histogram-Based Gradient Boosting puede manejar datos faltantes, esto genera que requiera el consumo de más recursos computacionales que podrían afectar la precisión de la predicción alcanzada por el mismo.

Para buscar alternativas más eficientes dentro de los modelos de Machine Learning se implementa el algoritmo no paramétrico KNN (K-Nearest Neighbors) para tratar los valores faltantes en el conjunto de datos. El objetivo de la imputación es reemplazar los valores faltantes por estimaciones basadas en otros valores observados. El método KNN se basa en la suposición de que los valores faltantes pueden ser estimados por los valores de observaciones similares (García, 2010).

El método KNN busca las k observaciones más cercanas a la observación con valores faltantes y toma un promedio ponderado de estos valores para estimar el valor faltante. La

ponderación se realiza en función de la distancia entre la observación faltante y las observaciones disponibles y el resultado final esperable es que ayude a mejorar significativamente la calidad de los modelos de aprendizaje automático.
Es importante destacar en este punto que la selección de la distancia a utilizar es fundamental en esta metodología y esta elección debe considerar el tipo de datos con el que se trabaja.

Para seleccionar el valor óptimo del parámetro K se consideraron dos caminos de búsqueda, el primero lo realizamos mediante una grilla probando diferentes valores y tomando como métrica RMSE, en todas las pruebas el valor elegido es 1, en cambio en la segunda opción directamente imputamos con valores de 1 a 7 vecinos para posteriormente probar diferentes modelos, los mejores resultados obtenidos se encontraban entre 4 y 5 vecinos. Por lo tanto, y teniendo en cuenta las diferentes ventajas y desventajas de elegir entre una u otra alternativa, el parámetro de vecinos final establecido para la imputación de valores nulos será de 5.

Luego de realizar la imputación de valores por KNN se ensayan diferentes modelos de regresión:

Se vuelve a ajustar el modelo HistGradientBoosting en esta segunda instancia sin la limitación de valores nulos. Los valores obtenidos se presentan en la Tabla 2:

Tabla 2: Resumen métricas Histogram-Based Gradient Boosting con imputación

Modelos	R2 ajustado	MAPE	MAE	RMSE
HistGradientBoosting sin imputación	0,27	13,76%	15,35	23,81
HistGradientBoosting	0,4	10,75%	12,21	21,57

Se logra mejorar de esta forma su desempeño disminuyendo la raíz del error cuadrático medio (RMSE) de 23.81 a 21.57 y aumenta el R^2 ajustado a un 0.4, sin embargo, es notorio que en esta instancia aún no contamos con suficiente poder explicativo.

A continuación, se ensaya un modelo de Árbol de Decisión (Decision Tree). Este modelo es una forma de árbol de decisión en el que cada nodo interno representa una característica (o atributo) y cada rama representa una posible salida para esa característica. Si bien entre sus ventajas se encuentra su fácil interpretación, que es capaz de manejar datos no lineales y que no requiere normalización, en la generalidad, su tendencia al sobreajuste y sensibilidad a la variación en los datos lo hacen menos adecuado para algunos tipos de problemas de regresión (Géron, 2019). Inicialmente este modelo evidenció un desempeño deficiente. Como esto puede deberse a la selección inadecuada de los valores de hiperparámetros, se decidió trabajar con una selección criteriosa de los mismos.

Del proceso de ajuste de los hiperparámetros, luego de un exhaustivo proceso de Grid Search, se decidió fijarlos en: máxima profundidad=5, y mínimo número de muestras para dividir un nodo interno=10. Con estos valores se logra una mejora sin embargo los valores aún no son estadísticamente aceptables:

Tabla 3: Resumen métricas Decision Tree Regressor con Grid Search

Modelos	R2 ajustado	MAPE	MAE	RMSE
HistGradientBoosting sin imputación	0,27	13,76%	15,35	23,81
HistGradientBoosting	0,4	10,75%	12,21	21,57
DecisionTreeRegressor + GS	0,25	12,56%	14,04	24,11

A continuación, se ensayó el modelo de regresión no paramétrico llamado KNeighbors Regressor, en el que se utiliza la distancia euclídea para calcular la proximidad entre observaciones y se estima mediante la media de los k vecinos más próximos o a menor distancia. Entre las ventajas principales de este método, se destacan su precisión y la sencillez de su implementación. En esta oportunidad el indicador de referencia alcanzó un valor de 0.41 mejorando el modelo de Hist Gradient Boosting, pero aún lejos de valores aceptables para los estándares de modelado, el cuadro se resume en la Tabla 4:

Tabla 4: Resumen métricas KNeighbors Regressor

Modelos	R2 ajustado	MAPE	MAE	RMSE
HistGradientBoosting sin imputación	0,27	13,76%	15,35	23,81
HistGradientBoosting	0,4	10,75%	12,21	21,57
DecisionTreeRegressor + GS	0,25	12,56%	14,04	24,11
KNeighborsRegressor	0,41	11,41%	13,10	21,51

A continuación, se evalúa la regresión por bosques aleatorios (RandomForest), que es una técnica de ensamble de modelos que combina múltiples árboles de decisión en una sola predicción y que gana robustez debido a la aleatoriedad en la selección de las características, además este algoritmo reduce la varianza del modelo (Breiman, 2001).

El único hiperparámetro modificado fue el número de árboles de decisión que se construirán en el modelo. En general, aumentar el número de árboles puede mejorar la precisión del modelo, ya que los resultados de las predicciones se promedian a través de múltiples árboles. Sin embargo, aumentar el número de árboles también puede aumentar el tiempo de entrenamiento del modelo y puede aumentar el riesgo de sobreajuste.
Para conjuntos de datos más pequeños, se suelen utilizar valores entre 10 y 100, por lo tanto, a fin de evitar el riesgo de sobreajuste utilizamos el valor mínimo sugerido de 10.

El regresor RandomForest aplicado con la previa imputación de valores mediante un KNN logró un coeficiente de determinación ajustado del 0.63 suponiendo una mejora

considerable con respecto al rendimiento de los modelos ensayados previamente, aunque consideramos que es posible optimizar aún más este valor. Al modelo citado le aplicamos un proceso de optimización de hiperparámetros mediante la técnica de Grid Search con el objetivo de encontrar la combinación óptima que maximice nuestras métricas de evaluación. Sin embargo, no se han logrado bajo esta metodología mejoras significativas en los siguientes coeficientes para evaluar la calidad de los modelos: R^2 ajustado, MAE y MAPE. El cuadro se presenta en la Tabla 5:

Tabla 5: Resumen métricas Random Forest Regressor con y sin Grid Search

Modelos	R2 ajustado	MAPE	MAE	RMSE
HistGradientBoosting sin imputación	0,27	13,76%	15,35	23,81
HistGradientBoosting	0,4	10,75%	12,21	21,57
DecisionTreeRegressor + GS	0,25	12,56%	14,04	24,11
KNeighborsRegressor	0,41	11,41%	13,10	21,51
RandomForestRegressor	0,63	9,27%	10,60	16,79
RandomForestRegressor + GS	0,63	9,68%	10,84	16,11

Finalmente, considerando sus ventajas, se evaluó el algoritmo de aprendizaje automático para la regresión que utiliza el método de Gradient Boosting de árboles de decisión llamado CatBoostRegressor.

Este modelo aplica la técnica de potenciación del gradiente para mejorar iterativamente la precisión de sus predicciones. En cada iteración, se agregan nuevos árboles de decisión al modelo y se ajustan a los residuos del modelo actual, lo que permite mejorar el ajuste del modelo a los datos (Li, 2017). Entre sus principales ventajas se encuentran su gran capacidad para manejar variables categóricas sin necesidad de transformarlas, procesamiento rápido y eficiente, regularización incorporada para prevenir el sobreajuste y mejorar la generalización del modelo, y en cuánto a sus desventajas podemos destacar un mayor tiempo de entrenamiento en comparación con otros algoritmos de regresión y la necesidad de configurar adecuadamente los hiperparámetros para alcanzar un buen rendimiento.

Con la implementación de este algoritmo se alcanzó un valor de R^2 ajustado del 0.70 indicando un ajuste razonablemente bueno del modelo. Los valores de las métricas resultantes se resumen en la Tabla 6:

Tabla 6: Resumen métricas CATBOOST Regressor

Modelos	R2 ajustado	MAPE	MAE	RMSE
HistGradientBoosting sin imputación	0,27	13,76%	15,35	23,81
HistGradientBoosting	0,4	10,75%	12,21	21,57
DecisionTreeRegressor + GS	0,25	12,56%	14,04	24,11
KNeighborsRegressor	0,41	11,41%	13,10	21,51
RandomForestRegressor	0,63	9,27%	10,60	16,79
RandomForestRegressor + GS	0,63	9,68%	10,84	16,11
CATBOOST	0,7	8,84%	9,90	15,15

Posteriormente se intentó optimizar mediante el ajuste de los hiperparámetros del modelo utilizando un Grid Search. Los valores probados en la grilla para cada uno de los hiperparámetros son los siguientes (detalles se pueden consultar en anexo 2):

Tabla 7: Resumen hiperparámetros (con marca los óptimos)

	Iterations	Learning_rate	Depth
Valor1	100	0,01	3
Valor2	200	✓ 0,1	4
Valor3	300	0,5	✓ 5
Valor4	✓ 400	0,9	7
Valor5			9

Luego de realizar la búsqueda exhaustiva de hiperparámetros, se han obtenido los siguientes resultados:

Tabla 8: Resumen métricas CATBOOST Regressor con Grid Search

Modelos	R2 ajustado	MAPE	MAE	RMSE
HistGradientBoosting sin imputación	✗ 0,27	✗ 13,76%	✗ 15,35	✗ 23,81
HistGradientBoosting	✗ 0,4	✗ 10,75%	✗ 12,21	✗ 21,57
DecisionTreeRegressor + GS	✗ 0,25	✗ 12,56%	✗ 14,04	✗ 24,11
KNeighborsRegressor	✗ 0,41	✗ 11,41%	✗ 13,10	✗ 21,51
RandomForestRegressor	✗ 0,63	✗ 9,27%	✗ 10,60	✗ 16,79
RandomForestRegressor + GS	✗ 0,63	✗ 9,68%	✗ 10,84	✗ 16,11
CATBOOST	✗ 0,7	✗ 8,84%	✗ 9,90	✗ 15,15
CATBOOST + GS	✓ 0,75	✓ 8,69%	✓ 9,65	✓ 13,85

Como conclusión, al obtener el coeficiente de determinación ajustado más alto, el MAPE, el MAE y el RMSE más bajos, se puede afirmar que el modelo CATBOOST Regressor con optimización de hiperparámetros mediante Grid Search es el más adecuado para explicar y predecir la variabilidad de la variable Glucemia. Estos resultados sugieren que este modelo logra una mayor precisión y capacidad explicativa en comparación con los demás modelos evaluados. Podemos comparar las métricas obtenidas de los siguientes modelos la Figura 9.

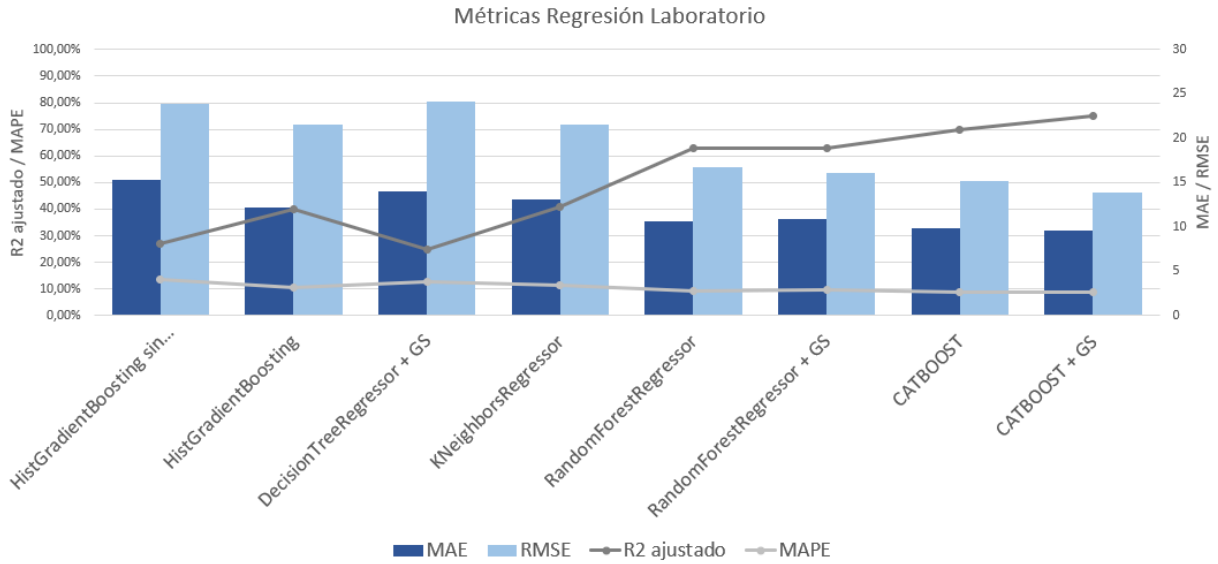


Figura 9: Evaluación de métricas sobre los modelos de Regresión para data set Laboratorio

La normalización de las métricas permite la comparación de múltiples indicadores en una misma escala, lo que facilita la visualización de las diferencias entre ellos. Una técnica comúnmente utilizada para la representación gráfica de métricas normalizadas es el gráfico radial. Este tipo de gráfico consiste en una representación gráfica en forma de rueda con múltiples ejes radiales que representan cada métrica. El valor de cada métrica se representa mediante un punto en el eje radial correspondiente, de modo que la distancia del punto al centro de la figura indica el valor de la métrica. De esta manera, el gráfico radial permite la comparación simultánea de múltiples métricas normalizadas y proporciona una visualización clara y concisa de las diferencias entre ellas.

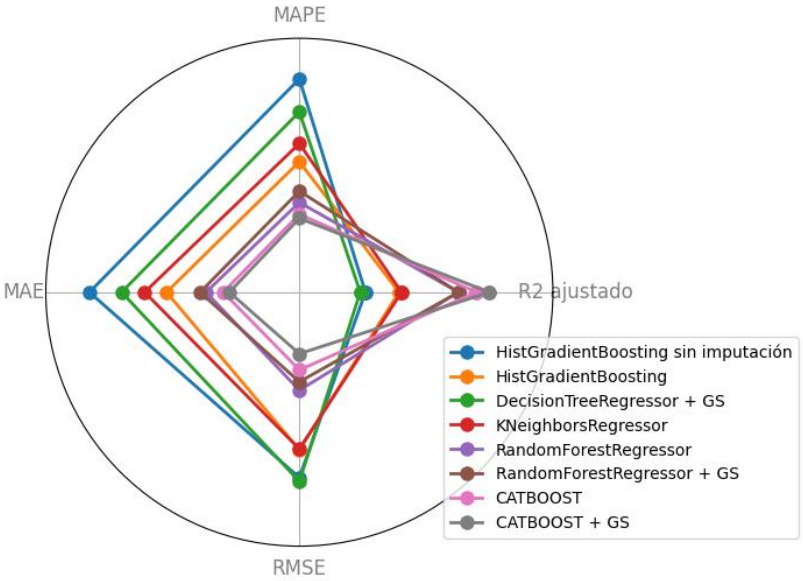


Figura 10: Rendimiento comparado de los modelos de Regresión para la base de Laboratorio

Identificamos las variables que presentan la mayor importancia en la predicción de la glucemia en nuestro modelo estrella de CatBoostRegressor, las cuales se presentan en la Figura 11:

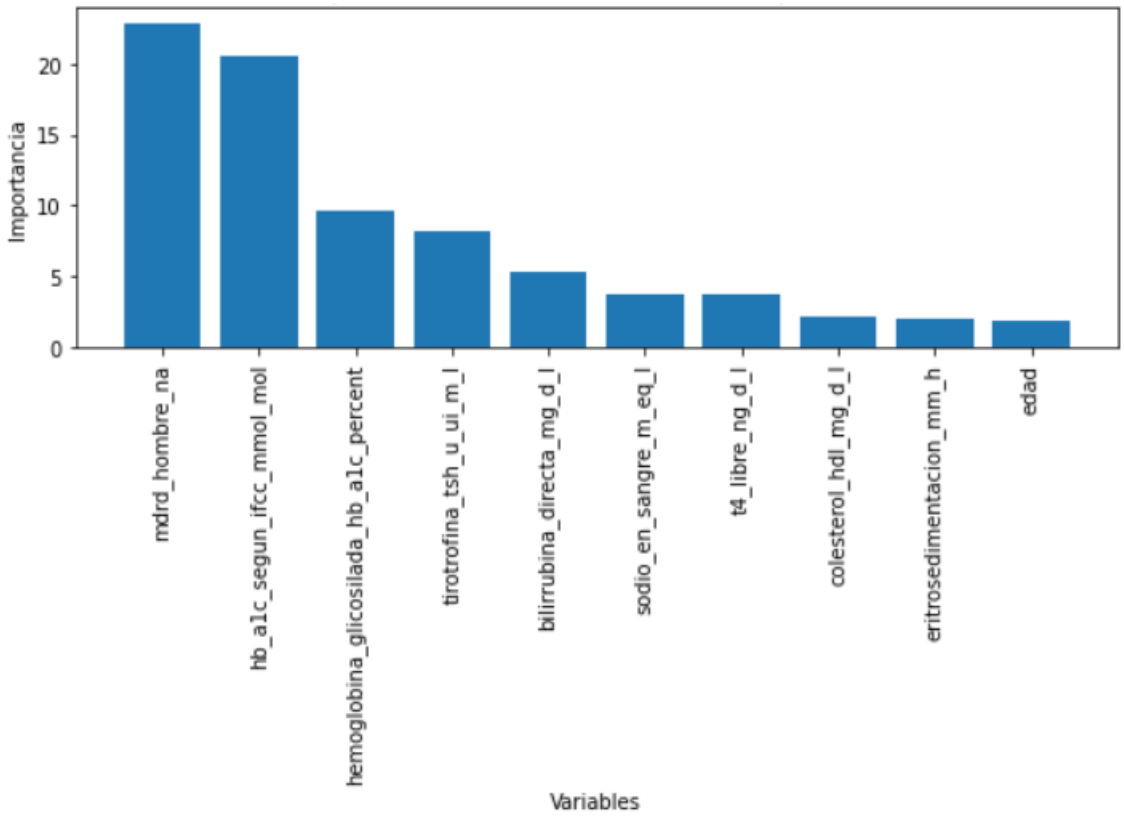


Figura 11: Las diez variables más representativas de la base de Laboratorio

3.3. Modelo de Predicción de Glucemia sobre Findrisc

Utilizamos el conjunto de datos de la encuesta Findrisc con el objetivo de realizar una estimación de los niveles de glucemia. Cabe destacar que esta base de datos supone cierta complejidad debido a su reducido número de registros. Para enfrentar esta complejidad, se llevan a cabo una serie de acciones que pretenden depurar el conjunto de variables. En este sentido, se procedió a eliminar aquellas variables que no ofrecen un valor predictivo y se desconsideró la variable "Colesterol Total", ya que después de aplicar diversas transformaciones se evidenció que más del 50% de los datos eran valores faltantes.

Las variables dummy son una técnica ampliamente utilizada en estadística y econometría, que permite convertir variables categóricas en variables numéricas binarias (con valores 1 o 0) (Hernández Sampieri, 2014). Sin embargo, la creación de variables dummy puede aumentar la complejidad del modelo y la multicolinealidad si existen muchas categorías en la variable original. En este sentido, evaluamos las ventajas y desventajas de esta técnica y se decidió aplicarla, ya que las variables a transformar presentan un número reducido de categorías, lo que nos permite aprovechar sus beneficios sin preocuparnos por sus posibles limitaciones.

Existen dos metodologías diferentes para generar variables dummy. La primera técnica consiste en crear una variable dummy para cada categoría de la variable categórica, excepto para una categoría que se utiliza como la categoría de referencia. La segunda técnica consiste en crear una variable dummy para cada categoría de la variable categórica, incluyendo la categoría de referencia. Ambas opciones presentan ventajas y desventajas. En el presente trabajo, se optó por la segunda opción, ya que es la más utilizada en análisis de varianzas. Sus principales ventajas son que reduce el riesgo de sesgo en los resultados del modelo y proporciona información detallada sobre las relaciones existentes, como contrapartida también tenemos que considerar que puede aumentar la complejidad del modelo.

Empleando la técnica anteriormente citada las siguientes variables se transformarán en binarias: Sexo, Actividad Física, Vegetales / Frutas, Hipertensión, Tratamiento HTA, Glucosa Si/No, Antecedentes Diabetes, Diabetes, Fumador.

En cuanto al tratamiento de valores faltantes, se emplea una técnica de imputación iterativa ampliamente utilizada en el aprendizaje automático conocida como iterative Imputer (Pedregosa, 2011). Este método utiliza modelos de regresión para estimar los valores faltantes a partir de los valores observados en otras variables, con el objetivo final de mejorar la precisión y validez de los análisis estadísticos y permitir la utilización de modelos que no admiten valores faltantes. Para aplicar este tipo de imputación se deben establecer

los valores de ciertos parámetros (ver anexo 3) que se pueden ajustar para obtener los resultados deseados.

Las pruebas realizadas evidenciaron que los mejores resultados se obtuvieron con los valores predeterminados, eso resulta razonable dado que estos valores han sido sugeridos por los desarrolladores en base a sus experiencias posteriores a la aplicación del algoritmo en diferentes conjuntos de datos. Se excluyeron los registros con valores nulos en la variable objetivo.

Además de las métricas ya presentadas se utilizarán otras métricas como el criterio de información Akaike (AIC), criterio de información Bayesiano (BIC) y la desviación mediana absoluta (MAD) para evaluar el rendimiento de los modelos.

El AIC (Akaike Information Criterion) es una métrica utilizada para evaluar la calidad de ajuste de diversos modelos estadísticos (Akaike, 1974). Este método se basa en la teoría de la información estadística, y se calcula a partir de la función de verosimilitud del modelo y el número de parámetros que se utilizan para ajustarlo. En términos generales, cuanto menor sea el valor del AIC, mayor será la calidad del modelo. Si bien considera un criterio de parsimonia y puede utilizarse en modelos lineales y no lineales, presenta desventajas, como la suposición de que el modelo es correcto, su limitación para evaluar la precisión absoluta del modelo, y su robustez depende del conjunto de datos y del modelo evaluado. En definitiva, el AIC es una métrica valiosa para evaluar la calidad de ajuste de modelos de predicción de variables continuas, aunque la utilizaremos junto con otras métricas debido a que su robustez se ve afectada cuando se utiliza en conjuntos de datos pequeños como es el caso del presente trabajo.

Su expresión es la siguiente:

$$AIC = -2 \ln L + 2k$$

Dónde:

- L es la función de verosimilitud del modelo
- k es el número de parámetros del modelo

Schwarz propuso el criterio BIC (Bayesian Information Criterion) como una medida para comparar modelos estadísticos (Schwarz, 1978). Aunque el BIC y el AIC buscan equilibrar la capacidad explicativa del modelo con su complejidad, difieren en la manera en que

penalizan la complejidad. Este criterio tiene sus raíces en el enfoque bayesiano de la estadística, se utiliza para comparar modelos alternativos en función de su probabilidad a posteriori, teniendo en cuenta la cantidad de parámetros del modelo y el tamaño de la muestra de datos. Si bien el BIC penaliza la complejidad del modelo con mayor fuerza que el AIC, lo que reduce aún más el riesgo de sobreajuste, tiende a seleccionar modelos más parsimoniosos que el AIC en conjunto de datos grandes. Su principal limitación es que tiende a seleccionar modelos más sencillos en conjuntos de datos pequeños en comparación con el AIC, además, el BIC asume normalidad en los errores de la regresión lo cual no es cierto en todos los casos.

Su expresión es la siguiente:

$$BIC = k * \ln n - 2 \ln L$$

Dónde:

L es la función de verosimilitud del modelo
k es el número de parámetros del modelo
n es el número de observaciones en los datos

También agregamos la métrica MAD (Mean Absolute Deviation, en español, Desviación Absoluta Media). Es una métrica comúnmente utilizada para evaluar la precisión de los modelos de predicción de variables continuas, pero en el presente trabajo utilizaremos su variante, en vez de tomar la media utilizaremos la Mediana (Median Absolute Deviation).
Es importante explicar los motivos que nos condujeron a esta opción; el primero es la robustez de esta medida definida como la mediana de los desvíos en valor absoluto relativos a la mediana, con esto nos referimos a su menor sensibilidad a residuales extremos que otras medidas (Mozumder, 2021).
La segunda es la distribución sesgada de los residuos, donde la media no resulta el parámetro de centralidad más conveniente. Por último, la tercera razón es la interpretabilidad de la medida y la sencillez de su computo.
En síntesis, la decisión se fundamenta en las características particulares de los datos de los que disponemos en este trabajo.
Su expresión es la siguiente:

$$MAD = Median(|Y_i - median(Y)|)$$

Dónde:

- |yi - ŷi| es el valor absoluto de la diferencia entre el valor real (yi) y el valor ajustado (ŷi)

Median es la mediana de valores absolutos de las diferencias definidas sobre el conjunto total de diferencias.

Ingeniería de Atributos

La ingeniería de atributos (Feature Engineering, FE) se define como el proceso de seleccionar, modificar y crear características o variables relevantes al modelo de predicción para incrementar su capacidad predictiva. Su principal propósito es preparar los datos de tal manera que el modelo pueda detectar patrones y relaciones que mejoren la precisión de la predicción (Kelleher, 2018).

En la literatura, se han propuesto diversas técnicas de ingeniería de características (FE), y en este estudio se optó por emplear el uso de entidades y relaciones para crear características complejas que podrían haber sido ignoradas en una revisión inicial de los datos, incluso por científicos de datos experimentados.
Sin embargo, este conjunto de herramientas debe ser cuidadosamente manejado a fin de evitar la construcción de características redundantes o irrelevantes.
En el proceso de construcción de estas variables se aplican técnicas de codificación de variables categóricas y generación de características a través de técnicas como el análisis de componentes principales (PCA) y análisis discriminante lineal (LDA).

No se observó una mejora significativa en la capacidad predictiva de los modelos mediante la aplicación PCA o LDA lo cual podría atribuirse a diversos factores, tales como la insuficiencia de datos o la presencia de datos de baja calidad y ese caso se evidencia en nuestro problema. También puede deberse a la presencia de sesgo en los datos o al desbalance de los datos. En este último caso existe la posibilidad de recurrir al submuestreo al sobremuestreo o a una combinación de ambas estrategias a fin de equilibrar los datos con los que se realiza el análisis.
Dado que no se alcanzaron mejores valores de desempeño como resultado de la aplicación de FE, se optó por avanzar en la construcción de los modelos prescindiendo de estas estrategias.

Estrategias de Validación

En relación con las estrategias de validación de modelos de aprendizaje automático, en lugar de utilizar el enfoque tradicional de dividir los datos en un conjunto de entrenamiento (train) y otro en un conjunto de validación o prueba (test) como se realizó para el caso de la base de datos de laboratorio se propone la técnica de la validación cruzada con k-fold.

El enfoque de validación de split en train y test es útil cuando se dispone de suficientes datos y se desea medir el rendimiento de un modelo en datos nunca vistos, se divide los datos en conjuntos de entrenamiento y se evalúa su rendimiento en el conjunto de prueba. Esto permite detectar problemas de sobreajuste y subajuste, como contrapartida, si la cantidad de datos disponibles para cada conjunto es pequeña, esta técnica puede resultar poco efectiva.
Por otro lado, cuando se tienen pocos registros, la técnica más recomendable es la validación cruzada con k-fold, ya que permite aprovechar al máximo los datos disponibles al generar varios conjuntos de entrenamiento y prueba (James, An Introduction to Statistical Learning: With Applications in R, 2013). Esto se hace dividiendo el conjunto de datos en k subconjuntos llamados folds. El modelo se entrena en k – 1 folds y se evalúa en el fold restante. Este proceso se repite k veces, de modo que cada fold es utilizado para evaluar el modelo una vez.
De esta manera, se pueden entrenar y evaluar el modelo varias veces, y utilizar la media de los resultados para obtener una estimación más robusta del rendimiento del modelo.

En resumen, y como consecuencia de que en el data set bajo análisis de Findrisc una vez realizada toda la depuración cuenta con una cantidad escaza de registros se considera utilizar el enfoque de validación cruzada con k-fold para asegurar una estimación precisa y robusta del rendimento del modelo.

Evaluación Comparativa de los distintos modelos

Se ensaya en primera instancia un modelo de regresión de mínimos cuadrados, con la dificultad de que los residuos de este no satisfacen los supuestos de normalidad ni homocedasticidad. Además, se distingue la presencia de multicolinealidad, razón por la cual se exploran modelos alternativos dentro del campo del machine learning.

El primer modelo de machine learning analizado es el Hist Gradient Boosting. Cabe destacar que durante la ejecución del método k-fold se llevó a cabo un proceso de búsqueda en la grilla de hiperparámetros con el fin de maximizar el rendimiento del modelo. Los resultados obtenidos arrojaron que el modelo explica apenas el 45 % de la variabilidad de la variable

de respuesta Glucemia. En la Tabla 4 se presentan las métricas correspondientes a este primer modelo.

Tabla 9: Resumen métricas Histogram-Based Boosting Regressor

Modelos	R2 ajustado	AIC	BIC	MAD
HistGradientBoosting + GS	0,45	1782	1903	7,55

Debido a que el desempeño predictivo del modelo seleccionado no alcanza las metas prefijadas anteriormente, se procederá a evaluar otro modelo de aprendizaje automático denominado Decision Tree Regressor. Dicha evaluación se llevará a cabo mediante la técnica de búsqueda en la grilla de hiperparámetros. Los resultados de este modelo se presentan en la Tabla 10.

Tabla 10: Resumen métricas Decision Tree Regressor

Modelos	R2 ajustado	AIC	BIC	MAD
HistGradientBoosting + GS	0,45	1782	1902,90	7,55
DecisionTreeRegressor + GS	0,38	1815	1936,21	7,14

Su desempeño en términos de AIC y BIC es inferior al del modelo anterior sin embargo lo aventaja en el MAD, lo que indica que tiene menor discrepancia entre sus valores ajustados y los reales.

El tercer modelo que evaluaremos es el Random Forest Regressor ajustando sus hiperparámetros mediante una grilla. El resultado de los ajustes se presenta en la Tabla 11.

Tabla 11: Resumen métricas Random Forest Regressor + GS

Modelos	R2 ajustado	AIC	BIC	MAD
HistGradientBoosting + GS	0,45	1782	1903	7,55
DecisionTreeRegressor + GS	0,38	1815	1936	7,14
RandomForestRegressor + GS	0,65	1652	1773	6,06

El modelo Random Forest Regressor con los hiperparámetros ajustados por Grid Search ha mostrado un desempeño superior en virtud de las métricas definidas; por lo tanto, bajo esta metodología, se explorarán dos métodos de optimización adicionales: la Optimización Bayesiana (OB) y la Optimización Aleatoria (OA).

La optimización Bayesiana es una técnica que se basa en la utilización de modelos probabilísticos para optimizar los hiperparámetros de un modelo de aprendizaje automático. La idea principal es construir un modelo probabilístico de la función objetivo, que en este caso es la métrica de evaluación que se desea maximizar o minimizar (por ejemplo, la precisión, o el error cuadrático medio). Este modelo se actualiza a medida que se van evaluando diferentes combinaciones de hiperparámetros y se utiliza para elegir la

siguiente combinación a evaluar, de manera que se vayan explorando las zonas más prometedoras del espacio de hiperparámetros (Brochu, 2010).
Por otro lado, la optimización aleatoria es una técnica que consiste en la evaluación aleatoria de combinaciones de hiperparámetros dentro de un espacio de búsqueda predefinido. Esta técnica es simple y fácil de implementar, pero puede requerir muchas evaluaciones de combinaciones de hiperparámetros para encontrar la más performante (Chen, 2016).

La optimización Bayesiana tiende a ser más efectiva que la optimización aleatoria para encontrar los mejores hiperparámetros de un modelo dado que utiliza una estrategia más inteligente y adaptativa al tomar en cuenta los resultados de las evaluaciones anteriores para guiar la búsqueda, pero también puede resultar riesgosa si la función objetivo (es decir, la métrica que se utiliza para evaluar el modelo) tiene múltiples óptimos locales, ya que esto podría desembocar en un análisis exhaustivo de una región que parece prometedora pero que en realidad no es la mejor, caso contrario a la optimización aleatoria que podría llegar a cubrir completamente el espacio de búsqueda para encontrar el mejor óptimo al no estar sesgado por suposiciones previas.

En general, si se dispone de suficientes recursos computacionales y se desea encontrar los mejores hiperparámetros de manera eficiente, la optimización Bayesiana es la mejor opción. En cambio, si el espacio de búsqueda es muy grande o complejo y se desea cubrir todo el espacio de manera eficiente, la optimización aleatoria emerge como la principal opción a tomar.

En base a lo expuesto las métricas obtenidas con ambos métodos de optimización lograron alcanzar los siguientes valores:

Tabla 12: Resumen métricas Random Forest Regressor con OA y OB

Modelos	R2 ajustado	AIC	BIC	MAD
HistGradientBoosting + GS	0,45	1782	1903	7,55
DecisionTreeRegressor + GS	0,38	1815	1936	7,14
RandomForestRegressor + GS	0,65	1652	1773	6,06
RandomForestRegressor + OA	0,61	1683	1804	4,81
RandomForestRegressor + OB	0,71	1600	1721	4,14

El siguiente modelo por evaluar es el Gradient Boosting Regressor, que es una técnica de ensamble que combina varios modelos de aprendizaje débiles y los ajusta a los datos de entrenamiento de forma iterativa (López, 2019). Su objetivo es minimizar la función de pérdida, que mide la diferencia entre las predicciones del modelo y los valores reales del conjunto de datos. En cada iteración, se agrega un modelo débil, que se ajusta a los residuos (diferencia entre las predicciones del modelo actual y los valores reales) del modelo

anterior. Para agregar el modelo débil, se utiliza un enfoque de árbol de decisión que divide los datos en subconjuntos más pequeños. El árbol de decisión se construye para minimizar la función de pérdida en los subconjuntos de datos. La combinación de varios modelos débiles permite mejorar la precisión del modelo y evitar el sobreajuste.

Su principal ventaja es su capacidad para producir modelos altamente precisos, lo que lo convierte en uno de los algoritmos más populares en la actualidad. Además, tiene la capacidad de manejar datos desequilibrados y que cuenten con una gran cantidad de ruido. Con respecto a sus desventajas puede ser susceptible al sobreajuste, especialmente si se utiliza con datos ruidosos o con una gran cantidad de características, presenta cierta lentitud de procesamiento y requiere más recursos computacionales en comparación con otros algoritmos de aprendizaje automático, especialmente cuando se utiliza con grandes conjuntos de datos o con modelos complejos.

La selección del modelo Gradient Boosting Regressor (GBR) como el modelo final se debe a que, en comparación con otros modelos evaluados, el Histogram-Based Gradient Boosting Regressor (HGBR) obtuvo el peor rendimiento. Este hecho despertó nuestra curiosidad y nos motivó a explorar un modelo que, aunque tenga características similares al HGBR, sea más adecuado para el conjunto de datos de la encuesta Findrisc.

El HGBR es un algoritmo que utiliza histogramas para acelerar el proceso de construcción del árbol de decisión y reducir el tiempo de entrenamiento. Esta técnica es útil para data sets con un gran número de registros, ya que permite reducir el tiempo de procesamiento y hacer que el algoritmo sea más eficiente. Sin embargo, cuando el data set es pequeño, los histogramas pueden no ser tan precisos y pueden producir sobreajuste.

Aunque el proceso de construcción de árboles de decisión de GBR puede ser más lento que el uso de histogramas, el GBR tiene una mayor capacidad para manejar data conjuntos con pocos registros y evitar el sobreajuste. Esto se debe a que el GBR puede ajustar de manera más precisa los modelos débiles a los datos de entrenamiento, lo que lo hace más adecuado para data sets pequeños.

Como conclusión y tomando en cuenta que la base sujeta a análisis no cuenta con una gran cantidad de registros el Gradient Boosting Regressor es el modelo seleccionado para el modelo final del presente trabajo obteniendo los siguientes valores con las diferentes técnicas de optimización de hiperparámetros:

Tabla 13: Resumen métricas Gradient Boosting Regressor con OA y OB

Modelos	R2 ajustado	AIC	BIC	MAD
HistGradientBoosting + GS	0,45	1782	1903	7,55
DecisionTreeRegressor + GS	0,38	1815	1936	7,14
RandomForestRegressor + GS	0,65	1652	1773	6,06
RandomForestRegressor + OA	0,61	1683	1804	4,81
RandomForestRegressor + OB	0,71	1600	1721	4,14
GBR + Grid Search (GBR + GS)	0,86	1704	1778	6,63
GBR + Optimización Aleatoria (GBR +OA)	0,74	1563	1684	5,56
GBR + Optimización Bayesiana (GBR +OB)	0,66	1640	1761	6,06

Al considerar el coeficiente de determinación ajustado como métrica base el Gradient Boosting Regressor con optimización de parámetros mediante Grid Search emerge como el mejor modelo para el trabajo con los siguientes hiperparámetros (detalles pueden ser consultados en anexo 4):

Tabla 14: Resumen hiperparámetros (con marca los óptimos)

	Learning_rate	Max_depth	Max_features	Min_samples_leaf	Min_samples_split	N_estimators
Valor1	0,01	2	None	1	2	50
Valor2	0,05	3	sqrt	2	3	100
Valor3	0,1	4	log2	3	4	150

Si bien vemos que el coeficiente de determinación ajustado es mejor en el modelo GBR + GS, tanto AIC y BIC logran mejores métricas en el segundo mejor modelo el GBR + OA.
Como ninguna de las métricas es la ideal, deben considerarse todas conjuntamente, además de tener en cuenta el contexto y el propósito del modelo. En el presente trabajo tomamos como métrica base el coeficiente de determinación ajustado debido a que se pretende obtener el modelo que mejor se ajuste a los datos, es decir, aquel que explique la mayor cantidad de variabilidad en los datos. Si bien ambos modelos arrojan valores muy similares para todas las métricas y teniendo como objetivo principal la métrica del coeficiente de determinación ajustado más alto, el modelo seleccionado sería el Gradient Boosting Regressor con optimización de hiperparámetros por Grid Search.

Al ver esta disparidad frente a las distintas métricas, se realizó un ensamble de estos dos modelos para evaluar la posibilidad de mejorar la capacidad predictiva, la realización de un ensamble de un mismo modelo con diferentes técnicas de optimización como el Gradient Boosting Regressor con Grid Search y el Gradient Boosting Regressor con optimización aleatoria puede resultar beneficioso en la búsqueda de la mejor combinación de hiperparámetros y mejorar la precisión predictiva del modelo final.
Combinar estos modelos puede ayudar a reducir el impacto de los valores subóptimos de los hiperparámetros y mejorar la precisión del modelo final.
Se utilizó la metodología de Voting Regressor, que combina múltiples modelos de regresión para producir una predicción única. Este tipo de ensamble se basa en el promedio de las

predicciones de los modelos base permitiendo reducir el riesgo de sobreajuste y mejorar la precisión general.

A pesar de que los modelos individuales obtuvieron resultados prometedores, la combinación de ambos modelos no logró mejorar significativamente la precisión predictiva. La falta de mejora significativa en el desempeño del modelo ensamblado puede tener su origen en varias razones, incluyendo el tamaño limitado del conjunto de datos utilizado, la falta de variabilidad en los datos y la selección inadecuada de los hiperparámetros y configuraciones de modelos individuales. Es importante tener en cuenta estas limitaciones y explorar diferentes estrategias de modelado para maximizar la precisión predictiva en conjuntos de datos limitados.

Gráficamente las métricas de los modelos analizados se reflejan de la siguiente manera (con MAD normalizado tomando como 100 % el mejor MAD obtenido):

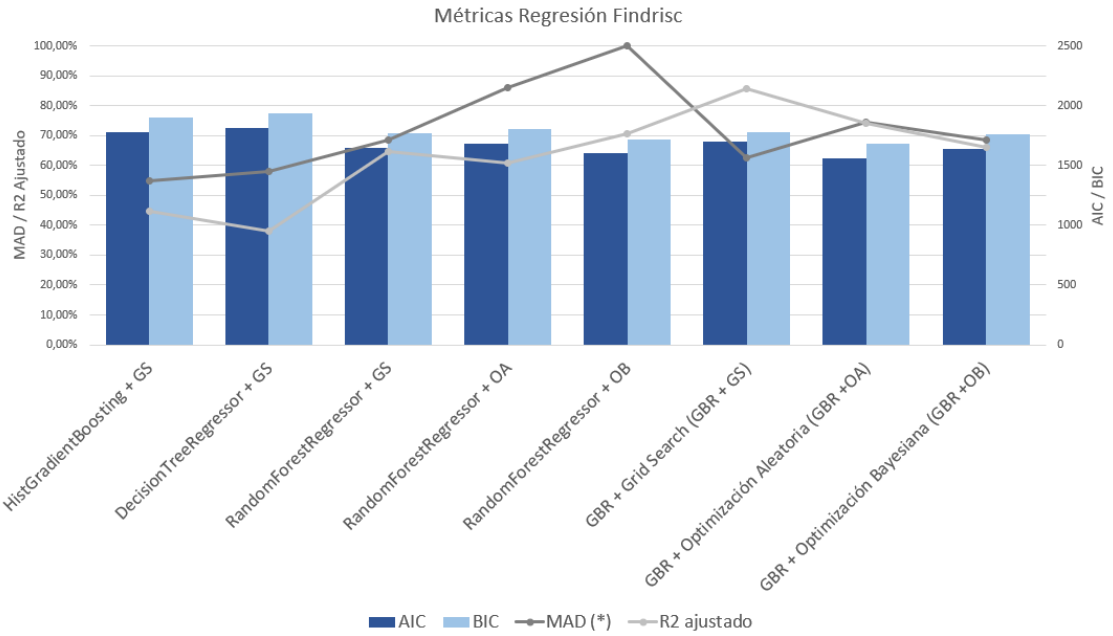


Figura 12: Evaluación de métricas sobre los modelos de Regresión para data set Findrisc con MAD normalizado

En ausencia de normalización de las métricas, se deben evaluar de forma independiente debido a que presentan distintas escalas. Para tal fin, generamos 4 gráficos representados en las Figuras 13, 14, 15 y 16.

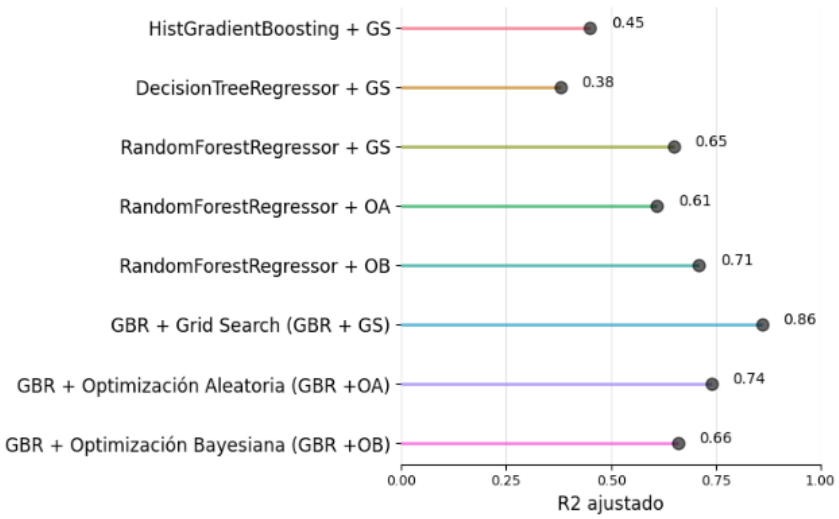


Figura 13: Evaluación comparativa de métricas R^2 ajustado

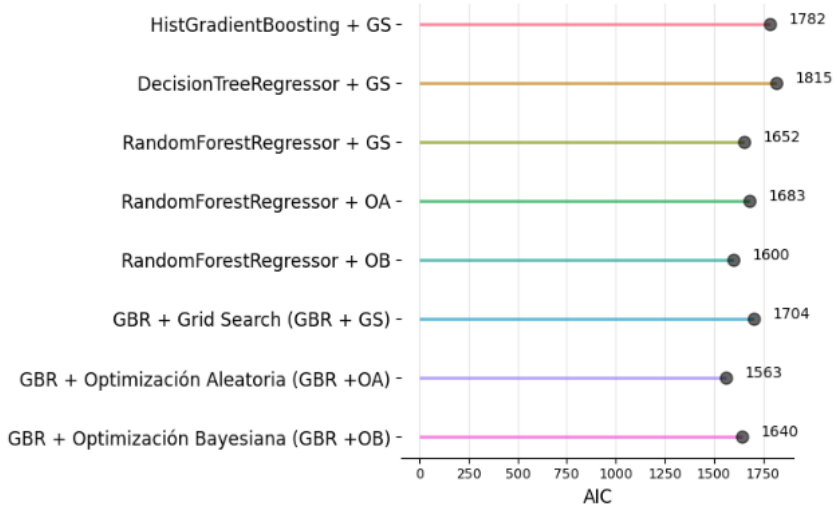


Figura 14: Evaluación comparativa de métricas AIC

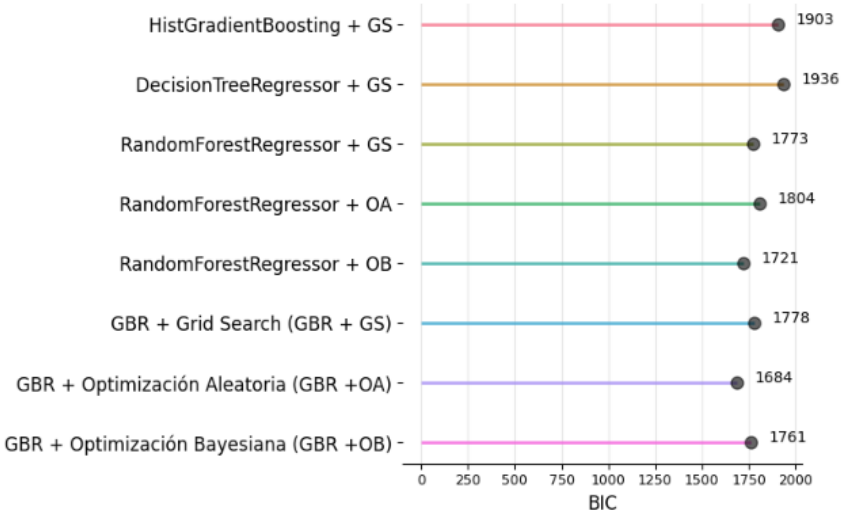


Figura 15: Evaluación comparativa de métricas BIC

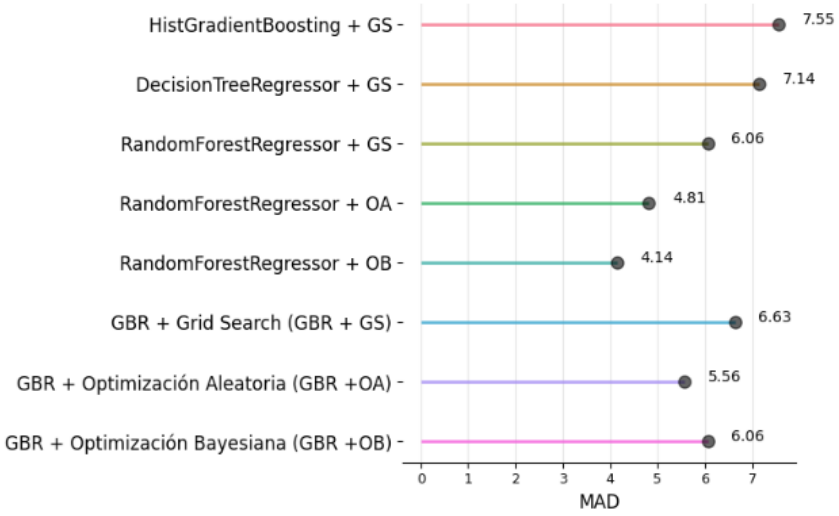


Figura 16: Evaluación comparativa de métricas MAD

Destacamos las siguientes variables como las de mayor importancia para la predicción de glucemia en nuestro modelo estrella de Gradient Boosting Regressor con optimización de hiperparámetros mediante Grid Search:

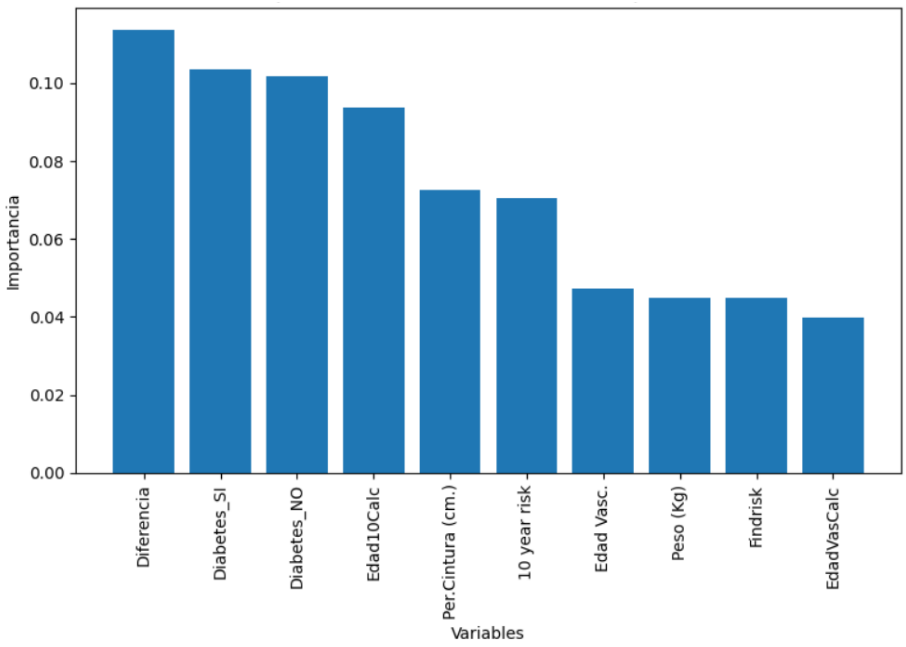


Figura 17: Las 10 variables más representativas para el data set de Findrisc

4. Capítulo 4: Conclusiones y trabajos futuros

En este trabajo se utilizaron diversas estrategias con el objetivo de abordar las dificultades derivadas de la escasez de registros y la alta incidencia de valores nulos presentes en ambos conjuntos de datos, generando estrategias que resultaran eficientes y que contribuyeran en la generación de modelos de predicción del valor de Glucemia tan precisos como fuera posible.

Asimismo, se aprovechó la información de la encuesta Findrisc, la información recogida en conjunto con la toma de la encuesta y se integró estos datos con los provenientes de análisis de laboratorio, donde algunos pacientes eran comunes a ambas bases.

En este camino se exploraron estrategias propias del campo médico para el análisis de la consistencia de la información y metodologías del aprendizaje automático y el análisis estadístico multivariado para detectar anomalías en las bases de datos disponibles.

Se presentó un conjunto de métricas para evaluar el desempeño de los modelos que se construyeran y se trabajó con la metodología de validación cruzada, que en una segunda etapa se enriqueció.

Inicialmente, los modelos propuestos en el ámbito de la estadística clásica, como la regresión, no lograban cumplir con los cuatro supuestos fundamentales requeridos para respaldar una inferencia robusta. Como respuesta a esta limitación, se produjo un giro hacia enfoque de modelado más contemporáneos, como las técnicas de aprendizaje automático. Los primeros modelos del campo del machine learning propuestos tampoco resultaron adecuados ya que no alcanzaban eficiencias aceptables en la precisión de sus predicciones, a pesar de que se propusieron una variedad amplia de estrategias de regresión del campo del machine learning.

Se trabajó entonces con ingeniería de atributos y esto condujo al descubrimiento de asociaciones interesantes entre la glucemia y otras variables disponibles en la base que en general permite descubrir patrones de asociación no necesariamente elementales o lineales. Sin embargo, estos descubrimientos interesantes para el campo de la salud no se tradujeron en mejoras significativas en relación con la precisión de las predicciones.

Para optimizar los hiperparámetros de los modelos se utilizaron diferentes algoritmos evaluando secuencialmente la mejora que cada uno de ellos aporta sobre el objetivo de predicción. En algunos casos el aporte de estas estrategias de optimización de los modelos

no condujo a mejoras significativas en ninguno de los criterios presentados para la evaluación del desempeño de los modelos.

Los resultados alcanzados, sin embargo, lograron mejoras altamente significativas respecto de los modelos propuestos inicialmente.

Considerando la métrica del coeficiente de determinación ajustado para el modelo de regresión basado en las variables de laboratorio se logró alcanzar un valor del 0.75 seleccionando el algoritmo CATBOOST Regressor mientras que en el caso de la base de Findrisc mediante el Gradient Boosting Regressor se alcanzó un valor de 0.86 para el coeficiente de determinación ajustado. Cabe destacar que en ambos casos la optimización de hiperparámetros se realizó mediante una búsqueda de grilla o Grid Search.

Estos resultados resultan alentadores en este campo del conocimiento porque además de tener un alto poder predictivo se convierten en una herramienta valiosa para estimar el valor de la glucemia en los pacientes.

Esta estimación es de bajo costo y el uso de estos modelos permite a los médicos estimar cuidadosamente el riesgo de padecer diabetes a diez años, pero también permite instrumentar estrategias que hagan posible demorar la declaración de esta enfermedad.

Asimismo esta herramienta también tiene valor para la OSEP en tanto le proporciona una estimación del riesgo de sus pacientes, le permite definir un subgrupo de pacientes que podrían estar cursando ya la enfermedad y no haber sido diagnosticados y fundamentalmente le permite realizar la previsiones económicas necesarias para afrontar el gasto que puede estimarse a partir del gasto histórico que esta patología insume y la estimación de la prevalencia de diabetes en la base de pacientes de la que dispone.

Si esta estrategia se extrapola a diversas instituciones de salud podría ser de valor para el sistema de salud nacional que podría, a partir de mejores y más tempranos diagnósticos, evitar un gran número de complicaciones derivadas del padecimiento de esta enfermedad.

En trabajos futuros se podrían extender estos resultados al conjunto de afiliados en su totalidad para poder construir una estimación adecuada del número de pacientes que padecen diabetes mellitus tipo 2, que podrían padecerla en los próximos años por estar en un estado prediabético y que tienen alta probabilidad de padecerla en función de sus registros actuales de glucemia y las respuestas recogidas por el cuestionario autoadministrado de Findrisc.

Además, podrían extenderse los resultados de este trabajo a la estimación de riesgos competitivos con el de diabetes entre los que podemos mencionar el deterioro de la función renal, el riesgo cardiovascular, las patologías oculares como la retinopatía diabética y la maculopatía diabética entre otras complicaciones frecuentes entre los pacientes diagnosticados con diabetes.

Estas estimaciones podrían realizarse basándose en bases de laboratorio similares a la que se analizó en este trabajo y complementarse con imágenes oftalmológicas que podrían revisarse mediante la aplicación de redes neuronales capaces de detectar el avance de la patología sobre la visión.

5. Referencias Bibliográficas

Akaike, H. (1974). *A new look at the statistical model identification.* IEEE Transactions on Automatic Control, 19(6), 716-723.

Alberti KG, e. a. (2007). *Diabetes UK. Diabetic Medicine.* 24, 451-463.

Ali, M. (April de 2020). *Pycaret: An open source, low-code machine learning library in Python*. Obtenido de https://www.pycaret.org

Armstrong, J. S. (2006). *Evaluating forecasting methods. Principles of forecasting: A handbook for researchers and practitioners.* 379-403.

Breiman, L. (2001). *Random forests. Machine learning.* 45(1), 5-32.

Brochu, E. C. (2010). *A tutorial on Bayesian optimization of expensive cost functions, with application to active user modeling and hierarchical reinforcement learning.* arXiv preprint arXiv:1012.2599.

Bundy, J. D.-J. (2020). *Cardiovascular health score and lifetime risk of cardiovascular disease: the cardiovascular lifetime risk pooling project.* Circulation: Cardiovascular Quality and Outcomes, 13(7), e006450.

Cardemil, F. (24 de Enero de 2017). *Análisis de comparación y aplicaciones del método de Bland-Altman: ¿concordancia o correlación?* Obtenido de https://www.medwave.cl/series/TyCEstadistica/6852.html

Chen, T. &. (2016). *Xgboost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.* (pp. 785-794).

Feskens, E., Virtanen, S., Räsänen, L., Tuomilehto, J., Stengård, J., Pekkanen, J., & Kromhout, D. (1995). *Dietary factors determining diabetes and impaired glucose tolerance: a 20-year follow-up of the Finnish and Dutch cohorts of the Seven Countries Study.* Diabetes care.

Friedman, J. H. (2001). *Greedy function approximation: A gradient boosting machine. Annals of statistics.* 29(5), 1189-1232.

Gagliardino, J. J. (2016). *Findrisc, una herramienta educativa. Revista de la Sociedad Argentina de diabetes.*

Gagliardino, J. J. (2016). *Prevención primaria de diabetes tipo 2 en Argentina: estudio piloto en la provincia de Buenos Aires. Revista argentina de endocrinología y metabolismo.* 53(4), 135-141.

García, S. F. (2010). *Advanced nonparametric tests for multiple comparisons in the design of experiments in computational intelligence and data mining: Experimental analysis of power.* Information Sciences, 180(10), 2044-2064. .

Géron, A. (2019). *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems.* O'Reilly Media.

Giavarina, D. (2015). *Understanding Bland Altman analysis.* Biochemia Medica.

Goodfellow, I. B. (2016). *Deep learning.* MIT Press.

Guzmán Rodríguez, S. F. (2016). *Estudio de detección del riesgo en diabetes en Atención Primaria según cuestionario FINDRISC en el Municipio de Gral. Pueyrredón (Estudio DR. DIAP).* Rev. Soc. Argent. Diabetes, 96-107.

Hastie, T. T. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* Springer.

Hastie, T. T. (2009). *The elements of statistical learning: Data mining, inference, and prediction (2nd ed.).* Springer.

Hernández Sampieri, R. F. (2014). *Metodología de la investigación (6a ed.).* McGraw Hill.
Hyndman, R. J. (2006). *Another look at measures of forecast accuracy. International Journal of Forecasting.* 22(4), 679-688.
International Diabetes Federation. (2009). *Diabetes Atlas.* 4.a.
International Diabetes Federation. (2013). IDF Diabetes Atlas (ed 6ta ed).
James, G. W. (2013). *An Introduction to Statistical Learning: With Applications in R.* Springer Science & Business Media.
James, G. W. (2013). *An introduction to statistical learning: With applications in R.* Springer.
Kelleher, J. D. (2018). *Data science: An introduction.* CRC Press.
Li, Z. W. (2017). *CatBoost: gradient boosting with categorical features support.* arXiv preprint arXiv:1706.09516.
Lindstrom, J., & Tuomilehto, J. (2003). *The diabetes risk score: a practical tool to predict type 2 diabetes risk* (Diabetes care, 26(3), 725-731. ed.).
López, V. (2019). *Gradient Boosting Regressor.* Springer International Publishing.
Ministerio de Salud de la Nación, Instituto Nacional de Esadísticas y Censos. (2015). Tercera Encuesta Nacional de factores de riesgo para enfermedades no transmisibles.
Morsanutto, A., Berto, P., Lopatriello, S., Gelisio, R., Voinovich, D., Cippo, P., & Mantovani, L. (2006). *Major complications have an impact on total annual medical cost of diabetes: results of a database analysis* (Journal of Diabetes and its Complications ed.).
Mozumder, M. S. (2021). *Performance Analysis of Machine Learning Algorithms on Data Mining and Analytics: A Comparative Study. Data Science and Engineering.* 6(1), 23-36.
Pedregosa, F. V. (2011). *Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research.* 12, 2825-2830.
Ruiz de Adana, S. C.-M. (2012). *Validación del FINDRISC (FINnish Diabetes Risk SCore) para la predicción del riesgo de diabetes tipo 2 en una población del sur de España.* Estudio Pizarra.
Saaristo, T., Peltonen, M., Lindström, J., Saarikoski, L., Sundvall, J., Eriksson, J., & Tuomilehto, J. (2005). *Cross-sectional evaluation of the Finnish Diabetes Risk Score: a tool to identify undetected type 2 diabetes, abnormal glucose tolerance and metabolic syndrome.* Diab Vasc Dis Res.
Schwarz, G. (1978). *Estimating the dimension of a model.* The Annals of Statistics, 6(2), 461-464.
Smith, J. D. (2018). *Comparing R-squared and adjusted R-squared: A simulation study. Journal of Applied Statistics.* 45(2), 346-356.
Whiting, D., Guariguata, L., Weil, C., & Shaw, J. (2011). *IDF diabetes atlas: global estimates of the prevalence of diabetes for 2011 and 2030.* Diabetes research and clinical practice.
Witten, I. H. (2016). *Data mining: Practical machine learning tools and techniques (4th ed.).* Morgan Kaufmann.
Zhang, Z. &. (2018). *Nonparametric Statistics: Theory and Methods.* Springer.

6. Índice de Figuras

7. Índice de Tablas

8. Anexos

8.1. Encuesta Findrisc

Control de salud diabetes

FINDRISK – con sólo 8 sencillas preguntas puede Ud. prever cuál es su riesgo de enfermar de diabetes tipo 2 en los próximos 10 años. ¡Aproveche esta oportunidad – realice esta prueba y permanezca sano durante el mayor tiempo posible!

Deutsche Diabetes-Stiftung

Qué edad tiene?

☐	Menos de 35 años	0 puntos
☐	De 35 a 44 años	1 punto
☐	De 45 a 54 años	2 puntos
☐	De 55 a 64 años	3 puntos
☐	Mayor de 64 años	4 puntos

Ha habido un diagnóstico de diabetes en, por lo menos, un miembro de su familia?

☐	No	0 puntos
☐	Sí, en mis parientes: abuelos, tíos y primos	3 puntos
☐	Sí, en mi familia directa: padres, hijos, hermanos	5 puntos

Qué perímetro de cintura tiene, medido a nivel del ombligo? (Si no tiene una cinta métrica, use un pedazo de cuerda y ayúdese con una regla)

	Mujeres	Hombres	
☐	Menos de 80 cm	Menos de 94 cm	0 puntos
☐	80 hasta 88 cm	94 hasta 102 cm	3 puntos
☐	Más de 88 cm	Más de102 cm	4 puntos

Tiene actividad física por lo menos 30 minutos diarios?

☐	Sí	0 puntos
☐	No	2 puntos

Con qué frecuencia come fruta, verduras o pan (de centeno o integral)?

☐	Diario	0 puntos
☐	No diariamente	1 punto

Le han recetado alguna vez medicamentos contra la hipertensión?

☐	No	0 puntos
☐	Sí	2 puntos

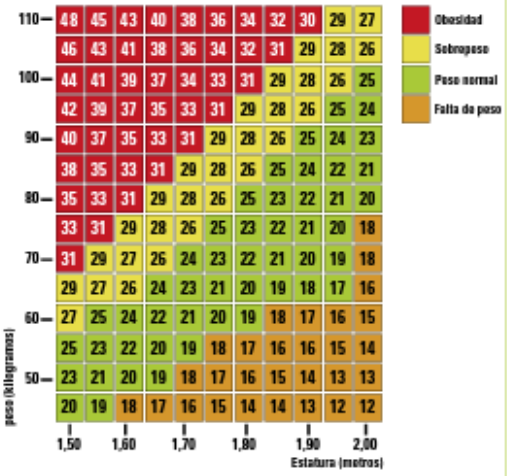
Le han detectato alguna vez, en un control médico, un nivel muy alto de glucosa (azúcar) en su sangre?

☐	No	0 puntos
☐	Sí	5 puntos

Cuál es la relación de su estatura y peso (Body-Mass-Index)?

☐	Menos de 25	0 puntos
☐	Entre 25 y 30	1 punto
☐	Más de 30	3 puntos

El índice de su masa corporal (BMI) lo calcula de la siguiente forma: Su peso (en kilogramos) dividido por su estatura (en metros) elevado al cuadrado (o simplemente según el cuadro, abajo)



puntos totales:_____

Deutsche Diabetes-Stiftung · Staffelseestraße 6 · 81477 München · www.diabetesstiftung.de

Control de salud diabetes

FINDRISK – su nivel de riesgo de contraer diabetes (en los próximos 10 años*)

Menos de 7 puntos — 1 por ciento*

Su nivel de riesgo es muy bajo. En su caso no es necesario un cuidado especial o de prevención. Sin embargo no estaría mal cuidar de su alimentación y realizar suficiente ejercicio.

De 7 a 11 puntos — 4 por ciento*

Para usted es recomendable un poco de cuidado, aunque el nivel de riesgo de contraer una diabetes no es muy alto. Si quiere ir sobre seguro, siga las siguientes reglas:

- En el caso de sobrepeso deberá intentar disminuir su peso en un 7 por ciento
- Manténgase en actividad, por lo menos, por media hora durante cinco días a la semana
- La grasa debería constituir,como máximo, sólo un 30 por ciento de su alimentación
- La parte de ácidos grasos no saturados (sobre todo en la grasa animal) no debería sobrepasar del 10 por ciento en su alimentación
- Consuma diariamente, por lo menos, 30 gramos de fibras vegetales (como las contenidas en productos integrales, verduras y frutas)

De 12 a 14 puntos — 17 por ciento*

Si usted se encuentra en este grupo de riesgo, no debería postergar, por ningún motivo, el tomar medidas preventivas. En este caso lo pueden ayudar consejos e instrucciones de expertos para cambiar su estilo de vida, los cuales puede aplicarlos usted mismo. Recurra a ayuda profesional si nota que de esa manera no se puede ayudar.

De 15 a 20 puntos — 33 por ciento*

Su nivel de riesgo es muy alto: una tercera parte de los pa cientes que corresponden a este grupo de riesgo contraen diabetes en los próximos 10 años. El subestimar esta situación puede traer graves consecuencias. Lo mejor sería recurrir a ayuda profesional. Haga una prueba de glucemia (azúcar en la sangre) en una farmacia y vaya a hacerse exámenes médicos (checkup a partir de los 35).

Más de 20 puntos — 50 por ciento*

Existe la necesidad de actuar inmediatamente, ya que es muy posible que usted ya sufra de diabetes. Eso pasa con el 35 por ciento de las personas que se encuentran sobre los 20 puntos. Una simple prueba de glucemia en su farmacia más cercana, por ejemplo, puede servir de ayuda como una información adicional. De todas formas, ésta no reemplaza un diagnóstico del laboratorio para descartar una diabetes ya existente. Por esta razón debería solicitar una consulta médica, inmediatamente.

*El riesgo en porcentaje = 4% significa, por ejemplo, que 4 de cien personas con este puntaje pueden contraer, en los próximos 10 años, una Diabetes Mellitus Tipo 2.

Usted puede disminuir el riesgo de la siguiente forma

Incluso pequeños cambios en su estilo de vida pueden apoyar su salud

Comer y beber de forma saludable

Más fruta y verdura	Todos sus alimentos deben contener, en lo posible, mucha fruta y verdura. Lo ideal es que aplaque su hambre solamente con estos alimentos.
Alimentos pobres en grasas	Al comprar productos lácteos elija las variantes con poca grasa. Coma carnes y embutidos magros pero con moderación. Por lo menos una vez por semana coma pescado.
Cocinar con poca grasa	Utilize sartenes con recubrimiento antiadherente, así se puede evitar el uso de mucho aceite al freir. En la cocina los aceites grasos deben ser sustituidos,principalmente, por el aceite de colza (al freir) y el aceite de oliva (en las ensaldas).
Bocadillos saludables	La comida rápida (fastfood) y la ya elaborada son bombas de calorías. Renuncie a ellas. La naturaleza le ofrece ricos productos para las comidas entre horas como: uvas, zanahorias o manzanas.
Bebidas saludables	Evite las bebidas que contienen azúcar. Aplaque su sed con agua mineral, zumos de frutas o té de hierbas.

Más ejercicio en su vida cotidiana

Tómese tiempo:	Haga ejercicio diariamente por 30 o 60 minutos. Elija actividades que pueda acomodar en su vida cotidiana.
Use el camino al trabajo como entrenamiento	Use, por ejemplo, el tiempo de espera en la parada de autobus y tense los glúteos y luego los músculos del estómago, alternativamente. Después balancéese sobre los dedos del pie, subiendo y bajando. Talvéz le alcance el tiempo incluso para ir a pie hasta la próxima parada.
Prefiera la bicicleta	Para hacer gestiones en las cercanías use la bicicleta. Colóquela en un lugar a su alcance, de tal manera que la pueda usar en cualquier momento y manténgala apta para funcionar.
Pruebe con la dinámica de grupo	Si le gusta hacer deporte en grupo, aprovéchelo. El establecer horas fijas para el deporte y el tener compañeros simpaticos puede ayudar a mantener la motivación en momentos de desánimo.
Los ejercicios correctos	Escoja tipos de deporte con una intensidad leve hasta media de esfuerzo. El Nordic-Walking, por ejemplo es un deporte ideal. Si usted suda levemente y puede conversar bién durante la práctica del deporte, entonces el esfuerzo que hace es el correcto.

Manténgase activo permanentemente

Fíjese objetivos realistas	Objetivos que no son fáciles de lograr, nos hacen tener mala conciencia y nos desmotivan
Introducir días de acción	De una a 3 veces por año debería crear condiciones claras, rompiéndo costumbres antiguas y ordenando su casa radicalmente. Esto vale para el refrigerador, así como para sótano y la sala. Cárguese de fuerza y energía para otros campos de la vida.
Engáñese a si mismo	Solamente las medidas que son fáciles de cumplir, pueden mantenerse en la vida cotidiana. Por ejemplo: las zapatillas de deporte que están en el corredor serán,probablemente, también usadas. Lo mismo sucede con el contenido del refrigerador: Los alimentos saludables colóquelos bien adelante ya que están más al alcance y son los primeros que se toman!

más información bajo:

- www.diabetes-risiko.de
- www.diabetesstiftung.org

Deutsche Diabetes-Stiftung

Deutsche Diabetes-Stiftung · Staffelseestraße 6 · 81477 München · www.diabetesstiftung.de

8.2. Hiperparámetros ajustados en CATBOOST Regressor mediante Grid Search

Iterations: es el número de iteraciones o ciclos de entrenamiento del modelo. Cada iteración implica la construcción de un árbol de decisión adicional y la actualización del modelo con los residuos de la iteración anterior. Afecta directamente el tiempo de entrenamiento y puede influir en su capacidad para ajustarse a los datos y generalizar en nuevos.

Learning_rate: la tasa de aprendizaje controla la cantidad en la que el modelo ajusta los residuos luego de cada iteración. Una tasa de aprendizaje alta puede permitir un ajuste rápido, pero también puede afectar la estabilidad del modelo y convertirlo en un modelo propenso a sobre ajustarse, en cambio, una tasa de aprendizaje baja puede llevar a un mejor rendimiento general, pero aumenta considerablemente el tiempo de entrenamiento.

Depth: La profundidad o tamaño máximo de cada árbol de decisión construido por el modelo. Afecta directamente la complejidad del modelo y su capacidad para capturar patrones en los datos. Un árbol con una profundidad excesiva puede permitir capturar patrones más complejos en los datos, pero también puede llevar a un sobreajuste y como consecuencia un rendimiento bajo en datos no vistos.

8.3. Hiperparámetros ajustados imputación iterativa

N_nearest_features: se refiere al número de características cercanas para utilizar en la estimación de los valores faltantes. Si no se lo establece, por defecto, toma todas las características disponibles.

Initial_strategy: estrategia de imputación inicial para los valores faltantes, por defecto el valor es la media, lo que implica que se utilizará la media de los valores existentes para completar los valores de datos faltantes.

Max_iter: número máximo de iteraciones para la imputación. El valor por defecto es 10.

Random_State: semilla aleatoria utilizada por el método

8.4. Hiperparámetros ajustados en Gradient Boosting Regressor mediante Grid Search

'learning_rate': El hiperparámetro de tasa de aprendizaje (learning_rate) controla la tasa a la que el modelo actualiza los coeficientes para ajustarse a los datos. Un valor bajo indica que el modelo se ajusta lentamente, mientras que un valor alto indica que el modelo se ajusta rápidamente. En este caso, el valor es 0.01 significa que cada vez que el modelo realiza una iteración para ajustar sus predicciones, los valores de los parámetros se actualizarán en un 1% del valor de la función de pérdida. Lo cual implica que el modelo se ajustará lentamente a los datos y puede requerir más iteraciones para alcanzar la convergencia.

'max_depth': El hiperparámetro de profundidad máxima del árbol (max_depth) indica cuántos nodos pueden haber en el árbol. Una profundidad alta puede permitir que el modelo se ajuste demasiado a los datos, mientras que una profundidad baja puede hacer que el modelo no se ajuste lo suficiente.

'max_features': El hiperparámetro de características máximas (max_features) indica la cantidad de características que se consideran al dividir cada nodo durante el proceso de construcción del árbol. Un valor bajo puede hacer que el modelo sea más generalizable, mientras que un valor alto puede aumentar la precisión del modelo.

'min_samples_leaf': El hiperparámetro de muestra mínima por hoja (min_samples_leaf) establece el número mínimo de muestras que deben estar presentes en cada hoja del árbol. Un valor bajo puede hacer que el modelo se ajuste demasiado a los datos, mientras que un valor alto puede limitar la capacidad del modelo para ajustarse a los datos.

'min_samples_split': El hiperparámetro de muestra mínima por división (min_samples_split) establece el número mínimo de muestras necesarias para dividir un nodo. Un valor bajo puede hacer que el modelo se ajuste demasiado a los datos, mientras que un valor alto puede limitar la capacidad del modelo para ajustarse a los datos.

'n_estimators': El hiperparámetro de número de estimadores (n_estimators) indica la cantidad de árboles de decisión que se deben utilizar en el modelo. Un valor alto puede aumentar la precisión del modelo, pero también puede hacer que el modelo sea más lento.