# Certamen N.º 1

Minería de datos

Integrantes: David Campos Muñoz,

Benjamín Montoya Oporto,

Nicolás Lagos Acevedo,

Javier Martínez Matamoros.

Docente: Isaac Daroch Ormeño

Fecha: 04-10-2025

**Índice**

[Certamen N.º 1 1](#_Toc210581951)

[**Sección A: Exploración (EDA) (15 min)** 3](#_Toc210581952)

[**Sección B Limpieza y Preprocesamiento (20 min + 20 min Pregunta 6)** 6](#_Toc210581953)

[**Sección C Minería de Datos (20 min)** 12](#_Toc210581954)

[**Sección Interpretación/Evaluación (15 min)** 15](#_Toc210581955)

## **Sección A: Exploración (EDA) (15 min)**

1. Formule una hipótesis medible para el dataset que le permite evaluar cuál es la variable dependiente y cuáles variables son más relevantes para la misma.

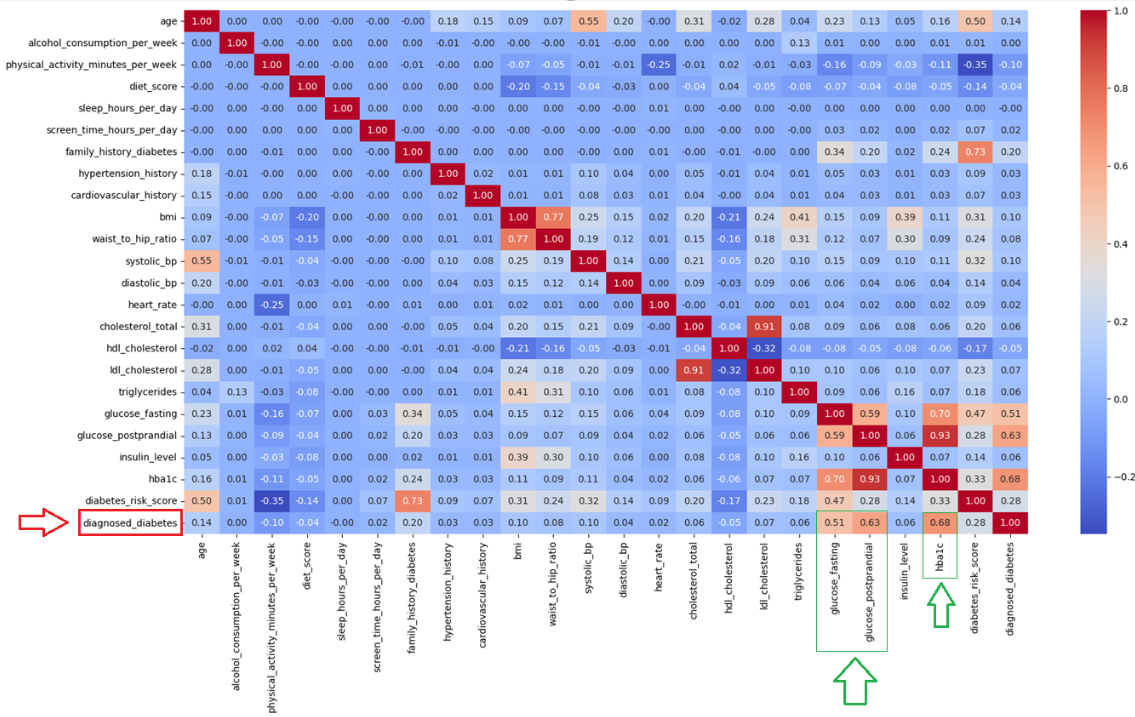
**Fundamente su respuesta**

RP: La variable dependiente es ***diagnosed\_diabete.*** *(En rojo)*

Las variables más relevantes son: *hba1c*, (correlación de 0.69), *glucose\_postprandial* (correlación de 0.64) y *glucose\_fasting* (correlación de 0.52). *(En verde)*

**Hipótesis:**

Se plantea que, a mayor hba1c, glucose\_postprandial y glucose\_fasting, mayor es la probabilidad de que diagnosed\_diabetes sea 1, o sea, positivo.



1. Indique en qué rango se concentra la mayor parte de los valores de una variable clave (de las que usted eligió) y qué significa para el problema.

RP: **En las variables claves:**

* **hba1c:** La mayoría de los datos se encuentran desde 5.97 hasta 7.07 en esta variable. Lo que significa que el 50% de los datos de pacientes diagnosticados, están al límite entre normalidad y prediabetes.
* **glucose\_postprandial:** La mayoría de los datos se encuentran desde 139 hasta 181, lo cual indica que mayormente los pacientes diagnosticados padecen de prediabetes.
* **glucose\_fasting:** La mayor parte de los datos se encuentran desde 102 hasta 120, lo cual indica que los niveles de azúcar en la sangre de los pacientes en ayuno son correspondientes a prediabetes.

1. Identifique las variables con mayor pérdida de calidad de datos (valores faltantes/nulos, ruido y duplicados). Liste las variables de menor a mayor porcentaje.

RP: En el DataFrame llamado ***diabetes\_dataset***, no se encontraron valores nulos ni duplicados, aunque existen algunos datos que pueden llegar a generar “ruido” como un bmi de 15 (bajo peso severo), 3 horas de sueño por día, 16.8 horas frente a una pantalla y una glucosa postprandial de 287. Por lo tanto, no se identifica una pérdida de calidad alta en los datos.

Imagen que contiene Tabla

El contenido generado por IA puede ser incorrecto.

1. Señale las variables en las que optaría por eliminar outliers (no imputar). Justifique con criterio técnico (IQR, STD, Media) y explique el impacto en el modelo/negocio.

En el caso de realizar eliminación de outliers en algunas variables, las que se escogerían para esto, serían aquellas que tiene valores imposibles de interpretar o que contengan información poco consistente en cuanto a su contexto, como, por ejemplo:

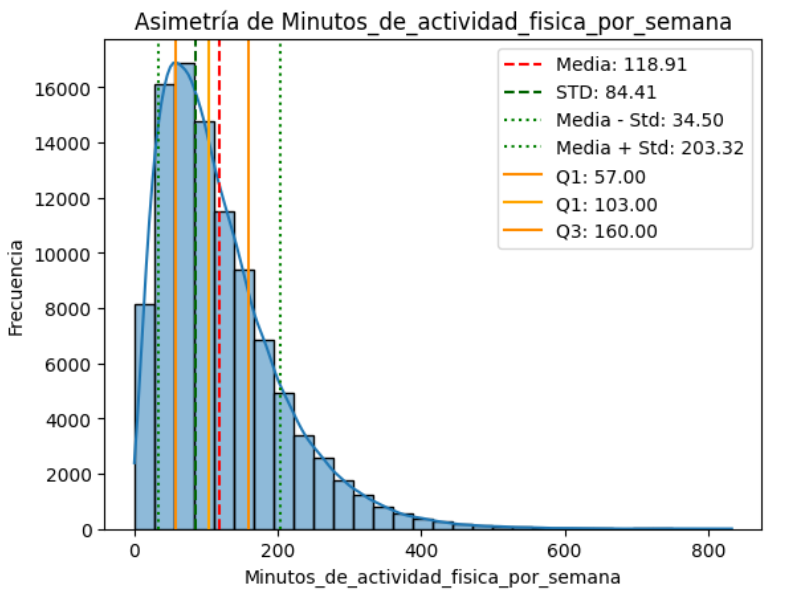
* ***diastolic\_bp:*** Menor a 40 o mayor a 120 mmHg
* ***cholesterol\_total:*** Menor a 100 o mayor a 400 mg/dL
* ***sleep\_hours\_per\_day:*** Menor a 3 o mayor a 15 horas
* ***bmi:*** Menor a 10 o mayor a 60

Estas variables entre estos valores representan outliers que no tendrían sentido en cuanto a comportamientos de la vida real, los cuales su eliminación no afectaría a la veracidad del análisis. Hay que señalar que, en el procesamiento, no se aplica eliminación de outliers, sino que se aplica winsorizar en k=3, que consiste en limitar un poco los valores extremos de los datos según el rango intercuartílico, evitando tener que perder información importante en el DataFrame.

1. Elija la variable con asimetría más alta. Grafique su histograma y consigne: media, desviación estándar, rango intercuartílico, luego interprete.

RP: El histograma de la variable Historial\_cardiovascular muestra una fuerte asimetría a la derecha (sesgo a la derecha), indicando que la mayoría de los pacientes realizan poca actividad física dentro de la semana, mientras que pocos presentan valores altos. La media es de 118.91 minutos/semana y la desviación estándar 84.41, con un rango intercuartílico de 103. Esto confirma que la mitad de los pacientes realiza entre 57 y 160 minutos de actividad física por semana.

De acuerdo con la hipótesis, esto se interpreta en que los pacientes suelen tener minutos de actividad física por debajo de la media.



## **Sección B Limpieza y Preprocesamiento (20 min + 20 min Pregunta 6)**

1. Explique su pipeline de imputación:
2. Detección y % de missing por variable;
3. Elección del método
4. Validación (comparar distribución antes/ después, cambio en media/IQR)
5. Variable más afectada y riesgos introducidos.

RP:

En este caso, con el DataFrame llamado ***diabetes\_dataset.csv,*** no presenta valores nulos o faltantes ***NaN*** en ninguna de las variables, por lo que no fue necesario crear un **pipeline de imputación.**

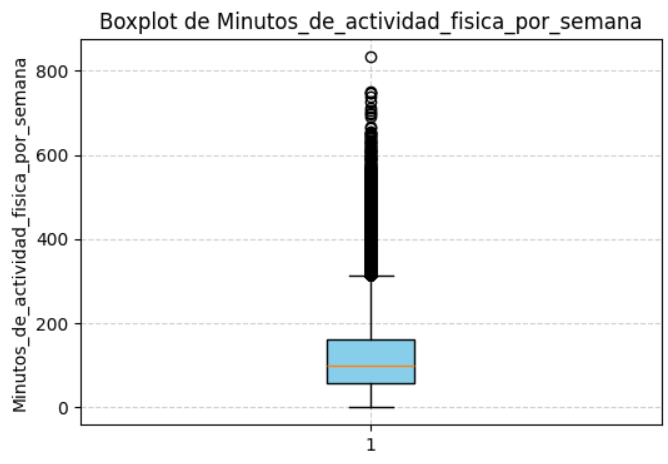
En contraste, al limpiar y preprocesar la data, se detectó bastantes outliers, los cuales afectan directamente la media y el análisis de la información. Variables como, por ejemplo: ***Glucosa\_en\_ayuno, Triglicéridos, Minutos\_de\_actividad\_física\_por\_semana y otros***.

Para ello, se aplicó winsorizar por k=3, para equilibrar la distribución de datos, evitando valores muy extremos de la data, posicionando la información en un rango más razonable para el análisis según el rango intercuartílico (IQR)

Este tratamiento de datos permitió mantener la cantidad total de los datos (100.000 registros) y equilibrar la distribución de estos sin eliminar información importante.

Al comparar la información antes y después del proceso de winsorizar, se observó que la media se mantiene correctamente, aunque el rango intercuartílico disminuye por el proceso dicho. Esto demuestra que variación de los datos se mantiene estable.

**Para acciones prácticas, vea el ejemplo a continuación:**

 Gráfico, Gráfico de cajas y bigotes

El contenido generado por IA puede ser incorrecto.

**Antes de winsorizar**

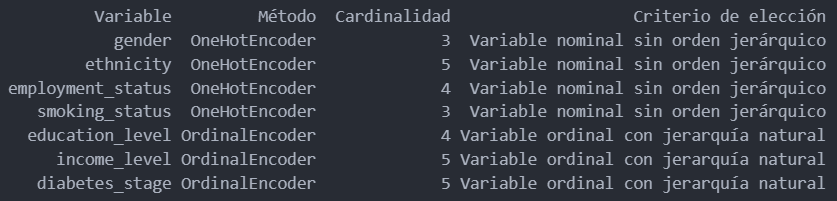
**Después de winsorizar**

El riesgo principal de este método es reducir de manera leve la variabilidad verdadera de los datos, aunque el beneficio que trae en consecuencia es seguir con la cantidad real de datos para el análisis.

1. Liste las variables codificadas y el método usado: para cada variable, indique si aplicó “OneHotEncoder” u “OrdinalEncoder”, la cardinalidad (# categorías) y el criterio de elección.

**RP:**

Durante el preprocesamiento de la data se identificaron las variables categóricas como: ***‘gender’, ‘ethnicity’, ‘education\_level’, ‘income\_level’, ‘employment\_status’, ‘smoking\_status’ y ‘diabetes\_stage’***. Para posteriormente revisar su cardinalidad y su método de codificación más conveniente según cada variable (ordinal o nominal):



1. Indique las variables con conversión de unidades o normalización equivalente, especificando unidad original à unidad final, fórmula de conversión, % de filas afectadas y motivo técnico (comparabilidad, coherencia física, interpretación del modelo)

RP:  
En el DataFrame llamado ***diabetes\_dataset.csv,*** lo que se hizo fue realizar un ajuste en las variables numéricas según el tipo de estas. Por ejemplo, en:

* **Variables clínicas**: Edad, BMI, la presión arterial, nivel de colesterol, entre otros.

A estas variables se transformaron sus diversas unidades de medida a valores estandarizados a través del proceso *StandardScaler*

* **Variables de estilo de vida**: Consumo de alcohol, minutos de actividad física, puntaje de dieta, horas de sueño y horas frente a la pantalla.

En estas variables, sus unidades de medida se transformaron a rangos entre 0 y 1, aplicando el proceso de *MinMaxScaler*

Esto con el objetivo de que todas estas variables estuviesen en una escala comparable, ya que, cada una de estas, estaban en unidades de medidas diferentes.

1. Liste las variables mal tipadas (número como texto, fecha como string, booleanos ‘SI/NO’ , IDs con ceros a la izquierda ). Para cada una, reporte: tipo actual, tipo correcto, regla de casting y validación (rango/regex).

RP:

Tras realizar un análisis en el Dataset no se visualizaron variables ‘mal tipadas’. Al realizar una revisión de los nombres y los tipos que tienen todas las columnas, se verificó que:

* No hay números guardados como texto, tanto en las variables de tema clínico como en las variables de estilo de vida.
* No existen fechas almacenadas como string en el Dataset.
* No hay booleanos escritos como SI/NO.
* No existen ID’s numéricos que requiera preservar ceros a la izquierda.

Debido a lo dicho anteriormente, no es necesario aplicar listas de variables mal tipadas ni aplicar reglas de conversión según la pregunta planteada, ya que todas las columnas se encuentran en el tipo de formato correcto.

1. Reporte el ratio de desbalance (IR = mayor/menor) de acuerdo con su variable objetivo, seleccione una técnica (RandomOver, RandomUnder) y justifique.

Tras analizar la variable objetivo llamada ***Diabete\_diagnosticada*** a través del proceso de análisis exploratorio de datos (EDA)*,* se observó la distribución de clases:

* Clase 1 (Con diabetes): 59.998 casos.
* Clase 0 (Sin diabetes): 40.002 casos.

Tras implementar el código que calcula el ratio de desbalance, se visualiza un resultado de 1.5, lo que indica un desbalance leve que no afecta de manera significativa el rendimiento del modelo. Por esta razón, no se aplicará ninguna técnica de balanceo de clases.

Sin embargo, en caso de tener que optar por balancear las clases, podríamos utilizar RandomUnderSampling para reducir la cantidad de datos de la clase mayoritaria, lo que también ayudaría en mejorar la velocidad del entrenamiento del modelo.

Texto

El contenido generado por IA puede ser incorrecto.

1. Reporte el número de componentes principales retenidos (K) y la varianza explicada acumulada (%). Incluya una tabla con PC\_i | varianza\_i | acumulada y explique qué implica para la reducción de dimensionalidad y la pérdida de información.

RP:

Se realizó el análisis de componentes principales PCA con el objetivo de reducir la dimensionalidad del conjunto existente del Dataset, conservando de manera objetiva la mayor parte de la información posible de la varianza total. En este análisis, se consideraron 15 variables numéricas del Dataset (excluyendo la variable objetivo llamada ***Diabete\_diagnosticada***).

Los componentes de PCA que se retuvieron para alcanzar una varianza explicada acumulada de un 95,57% fue de K=15.

En el gráfico ScreePlot, la curva de varianza acumulada supera justo el 95% de la proporción de varianza al incluir el PC15, desde este punto la pendiente en el gráfico tiende a aplanarse.

Gráfico, Gráfico de líneas

El contenido generado por IA puede ser incorrecto.

Así las cosas, al K=15, se retiene el 95,57% de la proporción de varianza total. Esto significa que el Dataset que resulta de este análisis preserva casi toda la información importante con solo 15 dimensiones de PCA, haciendo la información más simple y dejando una pérdida de información poco relevante.

**Tabla de resultados:**

**Imagen que contiene Interfaz de usuario gráfica

El contenido generado por IA puede ser incorrecto.**

## **Sección C Minería de Datos (20 min)**

1. ¿Qué división de datos utilizó? Escriba los porcentajes de train / validación / test.

RP:

Se uso una división de:

* **Entrenamiento:** 70,04% - 70.040 registros
* **Validación:** 14,96% - 14.960 registros
* **Test:** 15% - 15.000 registros

Se realizó por el motivo de mantener un equilibrio correcto entre el aprendizaje del modelo y la estimación de su desempeño en un entorno real.

Al dividir los datos en estas 3 categorías de porcentajes, lo que se busca es que el modelo tenga la suficiente información para aprender de manera correcta.

1. ¿Cuál fue el valor final de un hiperparámetro clave de su modelo (MLP: capas o alpha: Árbol: max\_depth; K-means) y por qué eligió ese valor?

RP:  
Se utilizó un modelo de red neuronal del tipo MLP (Perceptrón Multicapa), lo cual este modelo se aplica al conjunto de datos que se escalaron y procesaron anteriormente. De acuerdo con este caso, el hiperparámetro clave que fue ajustado fue **Alpha**. Este hiperparámetro le dice a la red neuronal cuanto debe controlarse o sujetarse para que no se memorice los datos netamente, sino que aprenda de patrones generales que influyan en una buena predicción del modelo.

Se probaron distintos valores en **Alpha**, como, por ejemplo: (0.0001, 0.001, 0.01 y 0.1). Finalmente, se eligió **Alpha=0.01** porque este valor ofreció más precisión en la validación de los datos.

1. Indique el tipo de problema y el modelo elegido: clasificación / regresión (predicción) / segmentación (clustering). Especifique variable objetivo/criterio, algoritmo, métricas de evaluación y por qué ese tipo calza con el objetivo analítico.

RP:

El problema que corresponde a este caso es de tipo clasificación binaria, ya que el objetivo del contexto del caso es determinar si una persona tiene o no diabetes. La variable objetivo que se utiliza en este caso es ***Diabete\_diagnosticada,*** la cual el valor 0 representa que no fue diagnosticado con diabetes y el valor 1 representa que sí fue diagnosticado con diabetes.

El modelo que fue elegido fue red neuronal MLP, debido a su capacidad de capturar relaciones entre los datos de acuerdo con las variables clínicas y de estilo de vida.

Para medir que tan bien funciono el modelo utilizado, se utilizó tres indicadores, los cuales son: ***Accuracy, recall y F1-score.*** Estos tipos de valores quieren decir que el modelo logro un grato equilibrio entre predecir correctamente y detectar la mayoría de los casos de personas con diabetes.

1. Con base en la métrica principal del modelo, indique si la evidencia apoya o refuta la hipótesis sobre la variable propuesta. Justifique con el valor de la métrica.

RP:  
La hipótesis planteada desde un inicio indica que, a mayor valor de *hba1c*, *glucose\_postprandial* y *glucose\_fasting*, mayor es la probabilidad de que una persona sea diagnosticada con diabetes. Con base en los resultados que entrego el modelo de red neuronal MLP, obtuvo un **recall** de 99% y un F1-score de 92,6%, lo que indica un rendimiento alto para identificar correctamente los casos positivos.

Así las cosas, a partir de esta evidencia, se puede dilucidar que la hipótesis propuesta es correcta, ya que el modelo logro predecir de manera precisa la condición de diabetes de los usuarios que, a partir de un aumento de *hba1c y las glucosas anteriores***,** estos factores están verdaderamente asociados con un mayor riesgo de diagnóstico positivo de diabetes.

1. Si el desempeño de su modelo es bajo, proponga un plan de mejora con acciones en datos, variables, preprocesamiento, modelo/hiperparámetros y evaluación. Indique qué cambia, por qué y cómo medirá la mejora.

RP:

En este contexto, el desempeño del modelo MLP no fue para nada bajo, ya que alcanzo métricas de **recall y F1-score** altas, llegando al 92,6% de exactitud. Esto da a entender que el modelo tuvo una buena capacidad para predecir correctamente los casos positivos de diabetes.

En contraste, si en un futuro existiera la posibilidad de que el rendimiento disminuyera, el plan de mejora a desplegar sería revisar y depurar varias etapas del proceso de modelo. La primera alternativa a considerar sería aumentar la calidad de datos, haciéndolo que sean más útiles para el entrenamiento, mediante la técnica de amputación de valores faltantes.

También seria de utilidad, desplegar técnicas de normalización de los datos, para que, en el proceso de realizar el modelo, el mismo pueda interpretar de manera mas exacta y trasparente las diversas variables y entregar un mejor resultado

## **Sección Interpretación/Evaluación (15 min)**

1. Indica un error típico que comete tu modelo (p. ej., confunde A con B / subestima valores altos / dos clústeres se mezclan) y qué harías para reducirlo.

RP:

El modelo tiende a cometer errores, principalmente subestimando casos positivos, es decir, predice que un usuario no tiene diabetes cuando realmente sí tiene diabetes (falsos negativos). Esto puede perjudicar que algunos pacientes en riesgo no sean identificados. Para reducir estos errores típicos se podrían afinar los hiperparámetros del MLP (capas, neuronas, Alpha) o equilibrar clases con RandomOverSampling / RandomUnderSampling.

1. Con ese desempeño, ¿lo usarías en la vida real? (SI/NO). Da una condición para usarlo o una mejora necesaria (p. ej., más datos, ajustar K, podar el árbol).

RP:

Como equipo, consideramos que este modelo Sí podría usarse en la vida real pero enfocado como una herramienta de apoyo y NO como un sistema de diagnóstico determinante. A pesar de que su rendimiento es bastante preciso, existe la posibilidad de que no sea 100% acertado según el tipo de contexto en que se despliegue.

Una mejora necesaria que encontramos como equipo, es que, para implementar realmente el modelo en un entorno real de predicción, sería necesario incluir más datos de diversos grupos de personas y de acuerdo con esto, realizar algunos ajustes al modelo de manera que su capacidad de detección sea más amplia. Si al realizar estos cambios, las pruebas del modelo se mantienen sólidas y estables, entonces se podría utilizar como una ayuda real para profesionales de salud enfocados en detección temprana de diabetes.

1. ¿Cuál fue el valor de la métrica de evaluación y, en una frase, eso es bueno, regular o bajo comparado con una referencia simple (azar/promedio/K al azar)?

RP:

El modelo desarrollado obtuvo **F1-score** de 92,6%, lo cual indica que tan acierta el modelo en los casos positivos y no equivocarse al realizar la predicción, esto representa un rendimiento alto.

Este valor obtenido es muy superior al que, por ejemplo, tiene un modelo de azar u otro modelo que solo predice que clase (K) es la más común. Estos modelos por lo general rondan en un **F1-score** de 50%.

Debido a esto, el rendimiento del modelo se considera de carácter bueno y confiable, ya que supera de manera acrecentada un modelo simple. Además, demuestra correctamente la predicción de diabetes en los pacientes.

1. Nombra una variable (o rasgo del clúster) importante y explica en una línea por qué crees que ayuda al resultado.

RP:

Una de las variables más importantes dentro del modelo desplegado es **Hba1c,** la cual representa el nivel promedio de glucosa en la sangre durante los últimos meses. Este factor es muy importante para identificar si una persona posee una cantidad insuficiente de azúcar en su cuerpo.

Esto permite al modelo notar con una alta precisión quienes tiene una mayor probabilidad de ser diagnosticados con diabetes.

1. ¿Si repites el entrenamiento con otra semilla o partición, el valor de tu métrica cambia poco o mucho? ¿Qué te dice eso sobre la estabilidad de tu modelo?

RP:

Al cambiar el entrenamiento con otra semilla (random\_state), el valor de la métrica cambia muy poco (diferencias de 0.1, 0.2), lo cual indica que el modelo es estable. Esto indica que el modelo produce resultados consistentes sin importar la semilla de entrenamiento.