Trabajo final



**MODELO PREDICTIVO PARA RIESGO CARDIOVASCULAR (RCV)**

**Proyecto integrador primer semestre - Maestría en Ciencia de datos y analítica**

**Autores**

**Equipo # 2**

Santiago Areiza Tamayo

Santiago Carvajal Torres

Pedro Gómez Bedoya

Lunes, 9 de junio de 2025

**Introducción**

Para una EPS o una IPS es muy importante identificar, prevenir y tratar a tiempo a los pacientes que tengan altas probabilidades de desarrollar enfermedades cardiovasculares, por lo cual es muy importante predecir el riesgo cardiovascular (RCV) de sus afiliados. Esto también va acorde con las iniciativas del gobierno nacional para los programas PyP (Programas de Promoción y Prevención en salud), por lo que un laboratorio clínico que oferte un producto de software que permita calcular dicho riesgo de manera confiable es un aporte valioso tanto para las finanzas del propio laboratorio como de los potenciales clientes (EPS e IPS) que pueden ahorrar mucho dinero en tratamientos a dichas enfermedades una vez ya fueron diagnosticadas.

Para abordar esta problemática, se propone el desarrollo de un modelo predictivo de riesgo cardiovascular, apoyado en técnicas de estadística, álgebra y analítica de datos. Desde la estadística, se aplicarán herramientas descriptivas y modelos probabilísticos que permitan cuantificar la relación entre variables clínicas y el riesgo cardiovascular. El álgebra será fundamental para la estructuración matricial de los datos y el desarrollo computacional de los algoritmos, mientras que la analítica permitirá la integración, transformación y análisis de grandes volúmenes de datos clínicos.

La solución se desarrollará mediante la construcción y validación de modelos predictivos, entre los cuales se consideran tanto enfoques tradicionales como métodos más avanzados:

* Métodos para entender y analizar holísticamente los datos, como *EDA*.
* Técnicas de reducción de dimensionalidad, como *PCA*.
* Modelos estadísticos clásicos, como la *regresión logística*, ampliamente utilizados por su solidez teórica e interpretabilidad en contextos clínicos.
* Modelos de clasificación usando *machine learning*, como *random forest*.

El modelo resultante será evaluado con métricas como exactitud, sensibilidad, especificidad, curva ROC y área bajo la curva (AUC), y se priorizará su capacidad de generalización y aplicabilidad en la práctica médica.

En conclusión, este trabajo busca desarrollar una herramienta predictiva útil para apoyar la toma de decisiones clínicas en la prevención de enfermedades cardiovasculares, contribuyendo así a una atención más oportuna, precisa y eficiente.

**Marco teórico**

El riesgo cardiovascular (RCV) es la probabilidad que tiene una persona de sufrir una enfermedad cardiaca o cerebral en los próximos 10 años. Se expresa en porcentaje. Existen varios modelos, Framingham, Score (Systematic Coronary Risk Evaluation), Regicor (Registre Gironi del Cor), Dorica (Dislipemia, Obesidad y RCV en España), desarrollados en diferentes países, para calcular de manera aproximada este riesgo, y tienen en cuenta los factores de riesgo más relevantes, como la edad, el género, la presión arterial, las cifras de colesterol, el consumo o no de tabaco y el hecho de sufrir o no diabetes. [1]

En el contexto colombiano, la detección temprana de este tipo de enfermedades cobran especial importancia ya que, como bien dicen Medias Benitez et al. en [1]: “*Puesto que las enfermedades cardiovasculares constituyen una de las principales causas de mortalidad y morbilidad en países desarrollados, es lógico que sea de gran interés el desarrollo de modelos de predicción del riesgo de padecer este tipo de enfermedades, tanto para intentar conocer los posibles mecanismos que afectan al aumento del riesgo, como para poder intervenir precozmente, mediante campañas preventivas, o en su momento con tratamientos terapéuticos…*”

En Colombia, el gobierno nacional ha estado promoviendo los programas de promoción y prevención en salud (PyP), los cuales son actividades basadas en el autocuidado y autocontrol, dirigidas a personas y familias para identificar riesgos y detectar tempranamente enfermedades y así brindar un tratamiento oportuno. [2]

Dichos programas buscan reducir la carga de enfermedades crónicas. Herramientas predictivas alineadas con estos programas permiten a EPS e IPS optimizar recursos, mejorar resultados clínicos y reducir costos. Un modelo predictivo confiable representa también un valor estratégico para laboratorios clínicos que deseen ofrecer soluciones innovadoras al sistema de salud.

Teniendo en cuenta la importancia de desarrollar un modelo predictivo con estas características, es de vital importancia usar la estadística para la creación y evaluación de dicho modelo, ya que proporciona herramientas para analizar y entender los datos clínicos, además de contar con métricas como exactitud, sensibilidad, especificidad, curva ROC y AUC, entre otras, que permiten evaluar el rendimiento de un modelo.

Los modelos predictivos permiten anticipar diagnósticos o eventos clínicos usando los datos disponibles. Para este caso particular, se utilizaron métodos y modelos de machine learning como PCA, regresión logística y random forest. El análisis de componentes principales, o PCA, es un método de reducción de dimensionalidad que a menudo se utiliza para reducir la dimensionalidad de grandes conjuntos de datos , transformando un conjunto grande de variables en uno más pequeño que aún contiene la mayor parte de la información del conjunto grande. [3]

La regresión logística es una técnica de análisis de datos que utiliza las matemáticas para encontrar las relaciones entre dos factores de datos. Luego, utiliza esta relación para predecir el valor de uno de esos factores basándose en el otro. Normalmente, la predicción tiene un número finito de resultados, como un sí o un no. [4]

Y por último, y no menos importante, el bosque aleatorio (o random forest) es un algoritmo de aprendizaje automático de uso común, registrado por Leo Breiman y Adele Cutler, que combina el resultado de múltiples árboles de decisión para llegar a un resultado único. Su facilidad de uso y flexibilidad han impulsado su adopción, ya que maneja problemas de clasificación y regresión. [5]

Por otra parte, la analítica de datos permite integrar y limpiar grandes volúmenes de datos provenientes de historias clínicas, laboratorios y registros médicos, asegurando así una alta calidad en el análisis. Para ello, es necesario contar con una metodología y una infraestructura robusta que permita asegurar que los datos sean extraídos, preparados, procesados y expuestos eficientemente. En este punto es necesario entender un concepto clave en la analítica de datos: el ciclo de vida de los datos.

El ciclo de vida de los datos es una secuencia de etapas por las que pasan los datos a lo largo de toda su vida útil. Los datos se separan en fases en función de diferentes criterios, y pasan por estas etapas a medida que completan diferentes tareas o cumplen ciertos requisitos. [6]

Dichas fases son:

1. **Generación o captura:** En esta fase, los datos llegan a una organización, normalmente a través de la adquisición desde una fuente externa o la recepción de señales, como los datos transmitidos por los sensores.
2. **Recolección de datos:** Durante la fase de recolección, se identifican, etiquetan y registran los datos de las fuentes potencialmente relevantes, ya que no todos los datos que se generan cada día se recogen o se utilizan.
3. **Mantenimiento y procesamiento de datos:** La etapa de mantenimiento y procesamiento del ciclo de vida de los datos es aquella en la que los datos se organizan y se procesan y luego se cuidan continuamente para mantenerlos accesibles y optimizados para los usuarios. En esta fase pueden someterse a procesos como la integración, la depuración y la extracción-transformación-carga.

Estos procesos incluyen la compresión de datos, en la que los datos se transforman en un formato que puede ser almacenado de forma más eficiente y el cifrado de datos, en el que los datos se traducen a otra forma de código para protegerlos de los problemas de privacidad.

1. **Uso de los datos:** Una vez que los datos se han procesado adecuadamente, llegan a la fase de utilización. En esta fase, los datos se utilizan para apoyar los objetivos y operaciones de la organización, y es el momento en el que se alinea el valor con la acción. Implica cómo se utilizan y se mueven los datos en la empresa.
2. **Intercambio de datos:** Durante esta fase, los datos se ponen a disposición de los usuarios de la empresa. Una vez que los datos están disponibles, pueden aprovecharse para una serie de análisis, desde el análisis exploratorio básico de datos y las visualizaciones de datos hasta técnicas avanzadas de minería de datos y aprendizaje automático.
3. **Archivo de datos:** El archivo de datos es la copia de los datos en un software para repositorios de investigación donde se almacenan por si se vuelven a necesitar en un entorno de producción activo. Luego, si es necesario, los datos archivados pueden restaurarse en un entorno de producción activo.  En esta fase, los datos se someten a un proceso de archivado que garantiza la utilización posterior.
4. **Destrucción de los datos:** En esta etapa final del ciclo de vida de los datos, se purgan de los registros y se destruyen de forma segura. Durante esta fase, los datos se eliminan de los archivos cuando superan el periodo de retención requerido o ya no tienen un propósito significativo para la organización. [6]

El producto de software resultante ayudará a la predicción y estratificación del riesgo de pacientes que puedan padecer enfermedades cardiovasculares en los próximos 5 años, de acuerdo con sus características físicas, hábitos de vida, condiciones preexistentes (antecedentes familiares) y sus historias clínicas (resultados de exámenes).

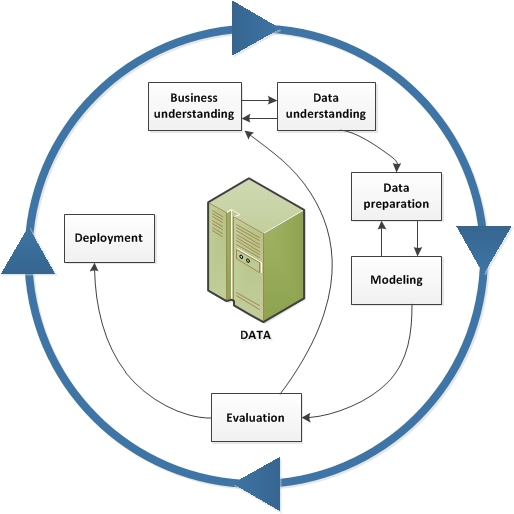
Esto beneficiará tanto a los pacientes, al ayudar a prevenir el riesgo de padecer enfermedades cardiovasculares y poder tratarlas a tiempo en caso de ser diagnosticadas, como a las EPS e IPS que se beneficien con los resultados entregados por el modelo resultante al desarrollar programas y estrategias para mantener y preservar adecuadamente la salud de sus afiliados.

Así mismo, debido a que el origen de los datos es de un laboratorio clínico real, dicho laboratorio se beneficiará económicamente al ofrecer y vender dicho modelo a sus clientes (EPS e IPS como se mencionó anteriormente) en caso de que el rendimiento y los resultados del producto final cumplan con sus expectativas para ofrecer dicho producto dentro de su oferta de servicios.

**Desarrollo metodológico**

**Metodología a emplear**

Para el desarrollo de este proyecto de analitica se usara la metodología CRISP-DM (Cross Industry Standard Process for Data Mining), reconocida en la industria por su enfoque estructurado, flexible e iterativo en proyectos de ciencia de datos. A continuación se muestra el ciclo de la metodología [7]:



Se listará cada ítem del ciclo de la metodología junto con su descripción asociada al proyecto:

1. **Comprensión del negocio:** El negocio se dedica a la realización de exámenes médicos de apoyo diagnóstico, incluyendo la toma de muestras y la entrega de resultados a pacientes.
2. **Comprensión de los datos:** Se explorarán los datos clínicos disponibles (edad, colesterol, presión, etc.), evaluando su calidad y relevancia estadística para el modelo del riesgo cardiovascular.
3. **Preparación de los datos:** Se realizará un análisis exploratorio de los datos clínicos disponibles para comprender su estructura y calidad, como paso previo a su uso en el modelado.
4. **Modelado:** Se entrenarán modelos de clasificación (como regresión logística o árboles de decisión) que permitan estimar la probabilidad de riesgo cardiovascular de forma precisa y útil para el contexto clínico.
5. **Evaluación:** Se validará el desempeño del modelo con métricas que den sentido al modelo de Machine Learning seleccionado y que genere tranquilidad con la realidad del proceso médico.
6. **Despliegue:** Aplicación en un ambiente tecnológico de la nube para facilitar el acceso, uso y escalabilidad del modelo predictivo.

**Entendimiento del problema**

El objetivo de este proyecto es desarrollar un modelo de machine learning que permita clasificar si un paciente tiene riesgo de desarrollar enfermedades cardiovasculares en los próximos 5 años de acuerdo con una serie de características propias de cada paciente como la edad, el peso, los antecedentes familiares, la presión arterial o los hábitos de vida (alimentación y actividad física). Este clasificador no entrega una probabilidad de RCV (riesgo cardiovascular) como tradicionalmente se conoce dicho indicador, si no que se trata de un clasificador que indica si un paciente activa cualquier alerta que determine si tendrá riesgo cardiovascular o no, es decir, si dicha probabilidad es mayor a 0%.

Sin embargo, un objetivo secundario de este proyecto es facilitar la visualización de los datos que permitan detectar patrones y establecer KPI’s que permitan a las empresas de salud analizar y tomar decisiones de acuerdo con los resultados obtenidos por el modelo. Es por esto que también se busca desarrollar un tablero en Power BI que sirva de ejemplo para que cualquier compañía pueda tener una idea de cómo utilizar los resultados entregados por el modelo desarrollado.

**Entendimiento de los datos**

El dataset “RCV.csv” fue construído a partir de una consulta en SQL Server en el servidor en la nube de Microsoft Azure donde están alojadas las diferentes bodegas de datos de los clientes de un laboratorio clínico. Por esta razón, fue necesario extraer e importar los datos a un archivo CSV mediante una consulta SQL para no tener que utilizar la infraestructura del laboratorio clínico y evitar problemas de permisos y accesos a la información al no pertenecer al directorio activo de la compañía.

Este dataset, en términos generales, cuenta con la información anonimizada de los pacientes del año 2024, con datos fisiológicos como la edad, el peso, el IMC, etc, también contiene los datos de antecedentes familiares como familiares que sufran de hipertensión o diabetes, datos de sus hábitos de vida como la ingesta de frutas y verduras y el sedentarismo y finalmente datos sobre su historia clínica como resultados de exámenes y diagnóstico de enfermedades.

**Variables:**

id\_persona

tipo\_sangre

rh

sexo

Etnia

Edad

Cie10Principal

Cie10Relacionado1

Cie10Relacionado2

Cie10Relacionado3

Cie10Relacionado4

Cie10Principal2

Cie10Relacionado4

Cie10Relacionado5

Cie10Relacionado6

HTA

DM

PAS

PAD

AlertaRiesgoHTA

IMC

IMCAnterior

Peso

PesoAnterior

G22-1

G22-2

953-1

C86-1

G06

AlertaDMSinFindrisk

PerimetroAbdominal

ActividadFisica

Alimentacion

antihipertensivos

TestFindrisc

InterpretacionFindrisc

AlertaTestFindrisc

AlertaConfirmadoHTAoDM

DifControlvsActual

AlertaHTAoDMSincontrolRec

AlertaHTANoControladoSinControlRec

H10

AlertaDMNoControladoSinControlRec

C34

AlertaHTAoDMLDLNoControladoSinControlRec

VigenciaC34

VigenciaH10

VigenciaD08

AlertaHTAoDMParaclinicos

PosibleAlertaHospitalizacion

ResultadoCreatininaActual

ResultadoCreatininaAnterior

TFGCyGActual

TFGCyGAnterior

AlertaTFGCyG

TFGActual

TFGAnterior

AlertaTFG

Albuminuria

Microalbuminuria

Proteinuria

AlertaAlbuMicroProteinuria

AlertaProgresionTFGCyG

AlertaProgresionTFG

PuntajeTotalSinFindrisc

Debido a que el origen de los datos son de clientes de un laboratorio clínico, estos datos son muy sensibles al tratarse de datos personales y confidenciales de los pacientes, por lo que es estrictamente necesario seguir los lineamientos del laboratorio clínico en cuanto a la forma de anonimizar los datos. Muy importante tener en cuenta que el archivo “RCV.csv” **NO** debe ser copiado ni distribuido por fuera de lo necesario para el desarrollo de este proyecto, ya que los datos siguen bajo la custodia del laboratorio clínico y esto podría violar sus normativas de gobernabilidad de datos.

**Preparación de los datos**

**Creación variable de respuesta**

Dado que el conjunto de datos no incluye explícitamente una variable que indique si una persona presenta riesgo cardiovascular, fue necesario construir una variable binaria denominada rcv (riesgo cardiovascular). Para ello, se consideraron siete variables tipo alerta, las cuales reflejan condiciones clínicas relevantes:

* AlertaRiesgoHTA
* AlertaDMSinFindrisk
* AlertaTestFindrisc
* AlertaConfirmadoHTAoDM
* AlertaTFGCyG
* AlertaTFG
* AlertaAlbuMicroProteinuria

Cada una de estas variables representa una condición de riesgo clínico. Por lo tanto, si una o más de estas alertas están activas (valor verdadero), se considera que la persona tiene riesgo cardiovascular (rcv = 1). En caso contrario, se clasifica como no en riesgo (rcv = 0).

#### **Depuración y validación de datos**

Durante la etapa de limpieza, se identificaron varias inconsistencias en los datos:

* Se encontraron valores de **índice de masa corporal (IMC)** superiores a 100, lo cual es fisiológicamente inviable. Estos registros fueron tratados como valores faltantes.
* En las variables de **presión arterial sistólica y diastólica**, se hallaron registros con valores de 0. Dado que una presión arterial de 0 es incompatible con la vida, estos datos también fueron marcados como faltantes.
* Se identificaron valores **negativos** en variables antropométricas como el **perímetro abdominal**, lo cual carece de sentido fisiológico. Estos registros también fueron tratados como faltantes.

Estas acciones fueron fundamentales para garantizar la calidad del conjunto de datos y evitar sesgos o errores en las etapas posteriores del análisis.

**Análisis Descriptivo**

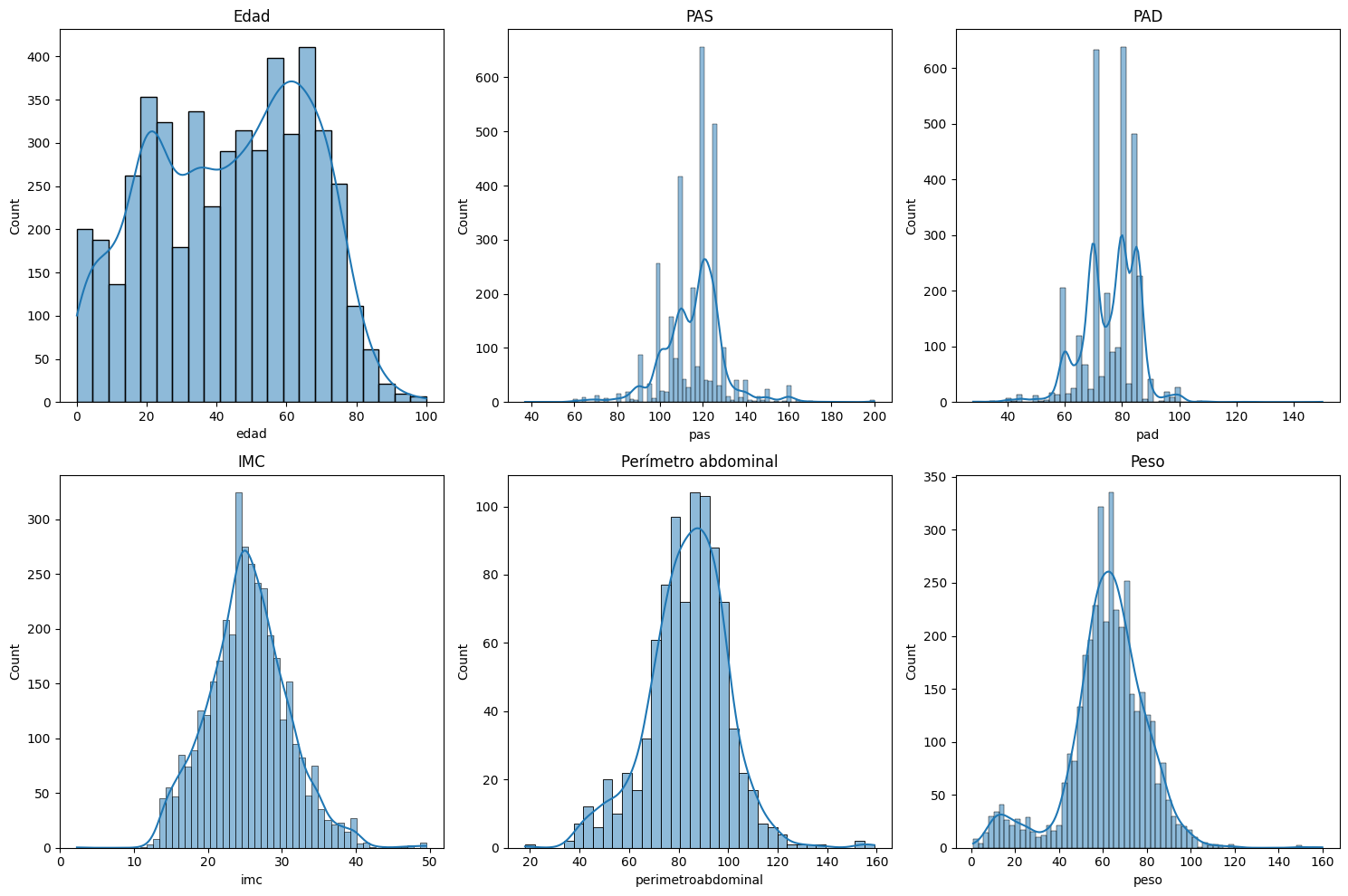
Este análisis tuvo como objetivo preparar y explorar un conjunto de datos clínicos para modelar el riesgo cardiovascular (RCV). A continuación, se presentan los principales hallazgos y decisiones tomadas durante el EDA:

1. **Depuración inicial:** Se eliminaron variables con más del 90% de valores nulos, entre ellas: 'c86-1', 'g22-1', 'g22-2', 'alimentacion', 'antihipertensivos', 'microalbuminuria', y 'proteinuria'. Esto permitió enfocar el análisis en variables con suficiente información para modelar.
2. **Matriz de Correlaciones - Reducción de la Redundancia:** Se calculó la matriz de correlaciones con el objetivo de evitar multicolinealidad y reducir la complejidad del modelo. A partir de este análisis, se identificaron variables con correlaciones superiores al 80%, las cuales fueron eliminadas por su redundancia. Estas variables fueron: pesoanterior, tfgcyganterior, resultadocreatininaanterior, pesoanterior y tfganterior.

Sin embargo, algunas variables altamente correlacionadas se mantuvieron debido a su alta relevancia clínica, validada por un experto del área. Estas incluyen: PAS (Presión Arterial Sistólica), PAD (Presión Arterial Diastólica), IMC, peso y perímetro abdominal.

1. **Selección de Variables para el Modelo:** Se definió un subconjunto de variables clave para el análisis predictivo, considerando aspectos demográficos, clínicos y de estilo de vida. Esta selección se basó en el análisis de correlaciones entre variables. Como resultado, se incluyeron las siguientes variables: id\_persona, sexo, edad, hta, dm, pas, pad, imc, albuminuria, rcv, tfgactual, peso, perimetroabdominal, actividadfisica, testfindrisc, resultadocreatininaactual y tfgcygactual.
2. **Análisis Descriptivo:** Se analizaron las distribuciones de seis variables clínicas continuas mediante histogramas con curvas de densidad (KDE), lo cual permitió identificar patrones poblacionales, simetrías y asimetrías relevantes:

* **Edad:** Presenta una distribución asimétrica hacia la derecha, con mayor concentración de pacientes entre los 30 y 65 años. Hay menos personas jóvenes y muy mayores.
* **PAS y PAD (presión arterial):** Ambas muestran una distribución bimodal, lo que sugiere la existencia de dos subgrupos: uno con presión normal y otro con presión alta.
* **IMC (índice de masa corporal):** Tiene una forma simétrica con leve sesgo a la derecha, centrada entre 25 y 30. La mayoría de los pacientes tiene sobrepeso.
* **Perímetro abdominal:** Su distribución es similar a una campana, con muchos pacientes cerca del valor medio (alrededor de 90).
* **Peso:** Muestra una forma simétrica pero con colas alargadas, indicando alta variabilidad en la población.



1. **Comparación de Grupos:** Se realizó una comparación de medias entre dos grupos de pacientes: aquellos sin riesgo cardiovascular (RCV = 0) y aquellos con riesgo cardiovascular (RCV = 1). Para ello, se aplicó la prueba t de Student para muestras independientes y se calculó el tamaño del efecto mediante el estadístico d de Cohen, con el fin de evaluar no solo la significancia estadística, sino también la magnitud de las diferencias observadas.

Se utilizó una significancia del 5% (α = 0.05). Dado que todos los p-valores obtenidos fueron inferiores a la significancia, se rechaza la hipótesis nula en todos los casos, concluyendo que existen diferencias estadísticamente significativas entre los grupos. Se llega a las siguientes conclusiones:

* **Edad:** El grupo con RCV tiene una media de edad significativamente mayor (59.6 años vs. 32.8 años), con un tamaño del efecto muy grande (d = -1.47).
* **Presión arterial (PAS y PAD):** Ambos valores son mayores en el grupo con RCV, con tamaños del efecto grandes (d ≈ -0.8 a -1.0).
* **IMC y perímetro abdominal:** Son más altos en el grupo con RCV, lo que indica mayor prevalencia de sobrepeso/obesidad (d ≈ -0.9).
* **Peso corporal:** También mayor en el grupo RCV, aunque con un tamaño del efecto algo menor (d = -0.76).
* **Creatinina sérica:** Ligeramente mayor en el grupo con RCV, con un efecto moderado (d = -0.4).
* **TFG (tasa de filtración glomerular):** Es significativamente menor en el grupo con RCV, lo que puede indicar función renal comprometida. El tamaño del efecto es grande y positivo (d = 0.87), porque una TFG más baja es peor clínicamente.

1. **Análisis de Variables Categóricas:** Se evaluó la relación entre variables categóricas y el riesgo cardiovascular (RCV) usando el test de Chi². Las variables de interés fueron: hipertensión (hta) y diabetes (dm).

Principales hallazgos:

* Las variables hta y dm presentaron una asociación estadísticamente significativa con RCV, con p-valores de 0.0000.
* En ambos casos, se observó una separación perfecta: todas las personas con hipertensión o diabetes pertenecen exclusivamente al grupo de alto riesgo (RCV = 1).
* Esta separación fue evidente en las tablas de contingencia:

Tabla hta vs rcv:

| hta | RCV=0 | RCV=1 |
| --- | --- | --- |
| 0 | 2959 | 1421 |
| 1 | 0 | 620 |

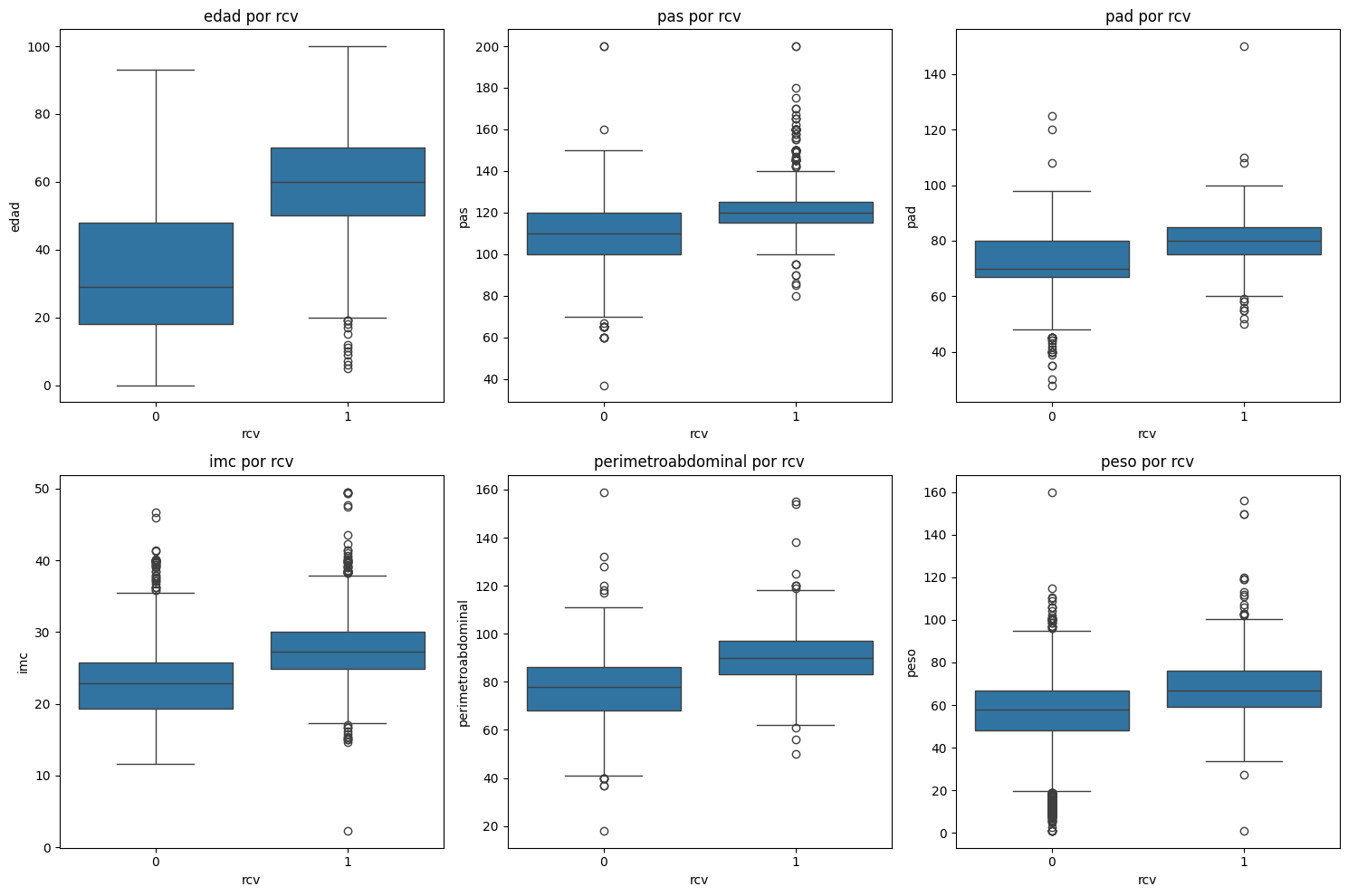
Tabla dm vs rcv:

| dm | RCV=0 | RCV=1 |
| --- | --- | --- |
| 0 | 2959 | 1806 |
| 1 | 0 | 235 |

Implicaciones para el modelado:

* La separación perfecta puede generar problemas en modelos como la regresión logística, donde los coeficientes pueden tender a infinito y dificultar la convergencia del modelo.
* Existe riesgo de sobreajuste, ya que el modelo puede aprender reglas demasiado estrictas (por ejemplo: "si hta = 1, entonces RCV = 1"), lo que reduce su capacidad de generalización.
* Modelos como árboles de decisión o random forest podrían asignar un peso desproporcionado a estas dos variables, ignorando otras que también aportan valor predictivo.

1. **Comparación Gráfica de Variables Clínicas por Grupo RCV:** Se utilizaron boxplots para visualizar la distribución de seis variables clínicas clave según el grupo de riesgo cardiovascular (RCV = 0 sin riesgo, RCV = 1 con riesgo). Esta representación permite identificar diferencias en la mediana, rango intercuartílico y presencia de valores extremos entre ambos grupos. Se muestra a continuación:



Principales observaciones:

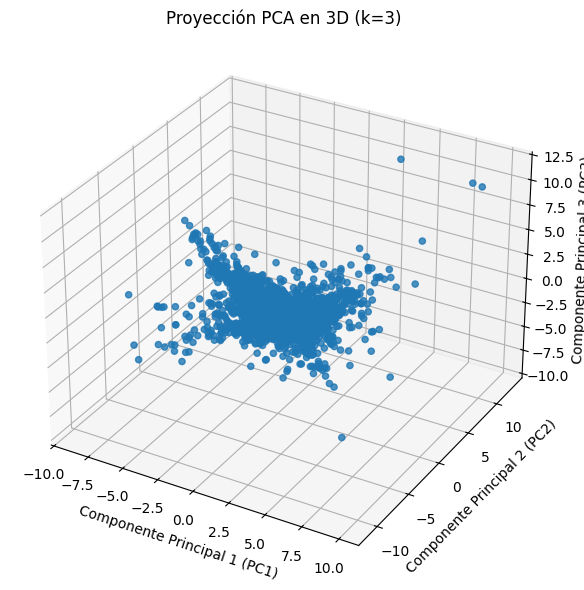
* **Edad:** El grupo con RCV tiene una mediana claramente superior. La mayoría de personas con bajo riesgo son jóvenes, mientras que las personas con RCV están concentradas en edades mayores.
* **PAS y PAD (presión arterial):** Ambos indicadores son visiblemente más altos en el grupo con RCV. Se observan múltiples valores extremos, especialmente elevados, lo que resalta su relevancia clínica en casos de hipertensión.
* **IMC y perímetro abdominal:** Estos indicadores de composición corporal también son mayores en el grupo con RCV. Las medianas y rangos intercuartílicos están desplazados hacia arriba, mostrando que el sobrepeso y obesidad abdominal son más comunes en personas con riesgo cardiovascular.
* **Peso:** Al igual que el IMC, el peso es más alto en promedio en el grupo con RCV y presenta mayor dispersión. La presencia de valores extremos indica que pacientes con mayor peso tienden a tener mayor riesgo.

**PCA**

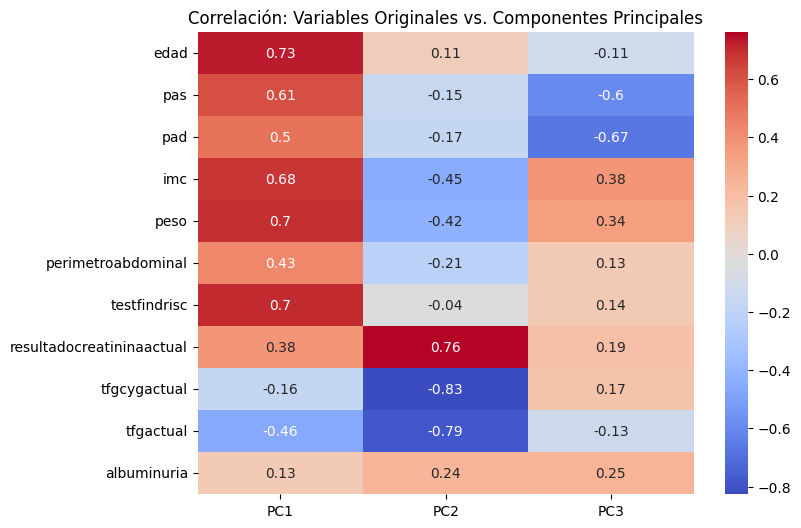
Este análisis aplica el método de Componentes Principales (PCA) para reducir la dimensionalidad de un conjunto de variables clínicas, manteniendo la mayor cantidad posible de información original. Las variables utilizadas en este análisis fueron previamente seleccionadas durante la etapa de análisis exploratorio de datos (EDA), específicamente en el paso de Matriz de Correlaciones – Reducción de la Redundancia, donde se filtraron atributos altamente correlacionados y se conservaron únicamente aquellas variables con valor clínico y estadístico relevante.

1. **Preparación de los datos:** Se cargaron los datos, seleccionando solo las columnas numéricas predictoras. Se imputaron valores faltantes con la mediana y se estandarizaron las variables para cumplir los requisitos del PCA (media cero y varianza unitaria).
2. **Etapas del PCA paso a paso:**

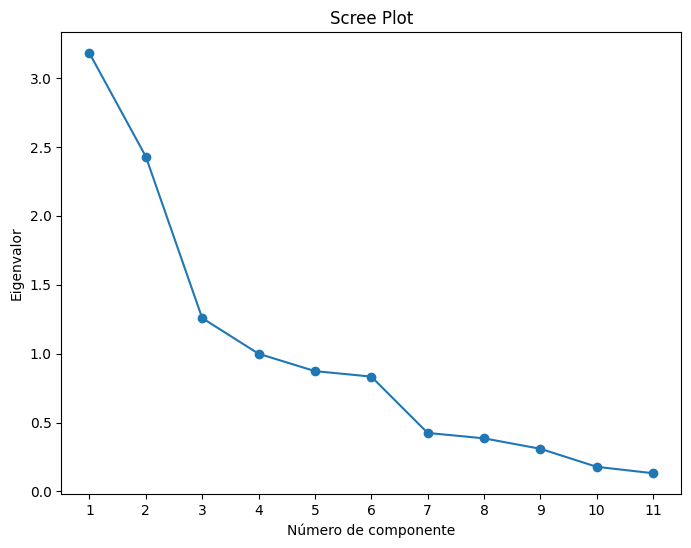
* Se centraron los datos y se calculó la matriz de covarianza (Σ).
* Se obtuvieron los autovalores y autovectores (eigenvalores y eigenvectores) de la matriz de covarianza.
* Se ordenaron los autovalores en orden descendente y se proyectaron los datos al nuevo espacio.



* Se construyó la matriz de correlación entre variables originales y componentes principales.

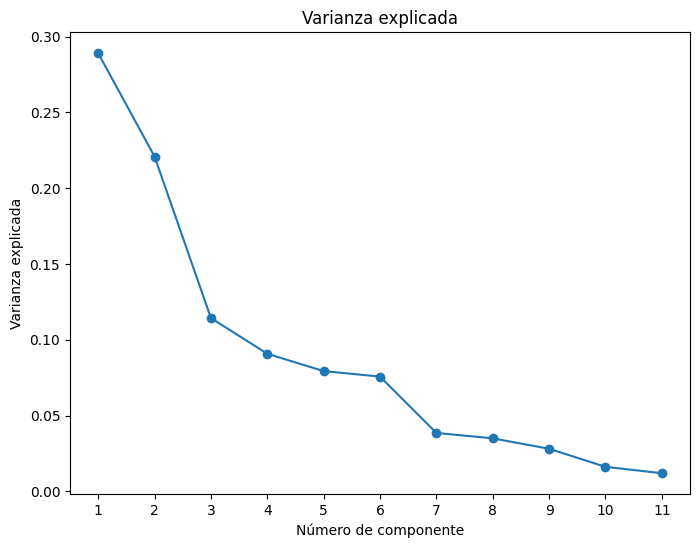


* Se graficó el scree plot para visualizar la magnitud de los autovalores.



1. **Selección del número de componentes:** Se aplicaron dos criterios:

* **Criterio de Kaiser:** Establece que se deben conservar solo autovalores mayores a 1, por lo que, sugiere conservar 3 componentes principales.
* **Varianza explicada acumulada:** Las 3 primeras componentes explican aproximadamente el 62.42% de la variabilidad total. Como se muestra en el siguiente scree plot:



1. **Interpretación de Componentes Principales:**

* **PC1:** Fuertemente correlacionada con edad (0.73), peso (0.70), test FINDRISC (0.70), IMC (0.68) y PAS (0.61). Refleja un patrón metabólico y de riesgo cardiovascular, ya que está fuertemente influenciada por la edad, el índice de masa corporal (IMC), el peso, la presión arterial sistólica (PAS) y el test de Findrisc. Estos son indicadores clásicos asociados a síndrome metabólico, obesidad y riesgo de enfermedades crónicas.
* **PC2:** Fuertemente correlacionada con tfgycgactual (–0.83), tfgactual (–0.79) y resultado creatinina actual (0.76). Representa el eje de función renal, tiene una fuerte correlación negativa con las tasas de filtración glomerular (TFG y TFGCYC), y positiva con la creatinina, lo cual es consistente con deterioro de la función renal. Un menor puntaje en PC2 sugiere peor función renal.
* **PC3:** Fuertemente correlacionada con PAD (–0.67) y PAS (–0.60). Componente relacionada con la presión arterial, tanto la diastólica (PAD) como la sistólica (PAS), y ambas de forma negativa. Esto significa que cuando PC3 es más bajo, la presión arterial tiende a ser más alta. Por eso, esta componente puede interpretarse como un indicador inverso de presión arterial: valores más bajos en PC3 están asociados con presión alta, y valores más altos con presión normal o baja.

**Análisis y conclusiones**

* El análisis de Componentes Principales (PCA) permitió reducir la dimensionalidad del conjunto de datos sin pérdida significativa de información, facilitando la visualización y comprensión de patrones clínicos complejos. Los tres componentes principales extraídos representan ejes fisiológicos bien definidos (riesgo metabólico, función renal y presión arterial) y pueden utilizarse como variables resumen en modelos predictivos.
* Además, cada componente puede ser interpretado como un indicador clínico de alerta temprana:
  + **PC1:** dominada por factores como edad, peso, IMC y test de riesgo metabólico, puede actuar como un indicador preventivo de riesgo cardiovascular y metabólico.
  + **PC2:** al capturar el deterioro de la función renal, puede interpretarse como un indicador temprano de daño renal crónico, útil para identificar pacientes con riesgo de progresión renal.
  + **PC3:** al estar relacionada con los niveles de presión arterial, puede interpretarse como un indicador temprano de alteraciones, útil para identificar personas con posible hipertensión o en riesgo de desarrollar presión alta.

**Modelos**

**Variables a utilizar**

Teniendo en cuenta los resultados del análisis descriptivo y el PCA las variables que utilizaremos para predecir “RCV” son:

1. PC1, PC2 y PC3
2. sexo
3. hta (hipertensión)
4. dm (diabetes)
5. actividadfisica

**Método de validación cruzada**

Para estimar de forma confiable el desempeño de cada modelo y evitar sesgos de partición, utilizamos **validación cruzada estratificada** con 5 pliegues. En cada iteración, el conjunto de datos se divide de modo que el 60% sirve para entrenar el modelo, el 20% sirve para testear el modelo y el último 20% para validar el modelo, manteniendo en cada pliegue la misma proporción de casos positivos de (rcv = 1) y (rcv = 0). Finalmente, se calcular las métricas (TP, FP, TN, FN, accuracy, precision, recall, F1, AUC y AP) en cada pliegue y se calcula su promedio y desviación estándar, lo que proporciona una estimación más robusta de la capacidad de generalización de los modelos.

**Regresión logística**

| **True Positives (TP)** | **190.6** |
| --- | --- |
| **False Positives (FP)** | **24.6** |
| **True Negatives (TN)** | **330.4** |
| **False Negatives (FN)** | **54.4** |
| **Accuracy** | **0.86** |
| **Precision** | **0.88** |
| **Recall** | **0.77** |
| **F1** | **0.82** |
| **AUC** | **0.93** |
| **Average Precision (AP)** | **0.91** |

Se observa un buen poder discriminativo con un AUC de 0.93, con excelente precisión de 0.89 y un equilibrio razonable entre sensibilidad y especificidad (0.77).

**Regresión logística regularizada (C = 0.01)**

Observamos que hta y dm tienden a dominar mucho el modelo, por lo que se decidió utilizar este método para evitar que dichas variables dominen el modelo y conduzcan a un sobreajuste.

| **True Positives (TP)** | **193.4 (+2.8 vs. normal)** |
| --- | --- |
| **False Positives (FP)** | **49.2 (+24.6)** |
| **True Negatives (TN)** | **305.8 (-24.6)** |
| **False Negatives (FN)** | **51.6 (-2.8)** |
| **Accuracy** | **0.832 (–0.036)** |
| **Precision** | **0.798 (–0.088)** |
| **Recall** | **0.789 (+0.011)** |
| **F1** | **0.793 (–0.035)** |
| **AUC** | **0.911 (–0.019)** |
| **Average Precision (AP)** | **0.879 (–0.035)** |

Con estos resultados podemos observar que una regulación “fuerte” aumenta ligeramente la sensibilidad (Recall), sacrificando parte de la precisión y la exactitud global.

Los coeficientes de hta y dm quedan mucho más cerca de cero, reduciendo su influencia, lo que mitiga el riesgo de que sujetos con hipertensión o diabetes sean “sobredeterminados” por el modelo.

Es necesario recalcar que aún así penalizando las dos variables que explican más el rcv, aún así el modelo es capaz de ser bastante decente.

**Random Forest**

Siguiendo la recomendación del profesor de Estadística, decidimos también meter un modelo de machine learning más robusto como puede ser random forest. Para la siguiente tabla, lo comparemos con el modelo de regresión logística normal ya que en general tuvo mejor rendimiento que el regularizado.

| **True Positives (TP)** | **213.6 (+23.0 vs. normal)** |
| --- | --- |
| **False Positives (FP)** | **24.6 (mismo que normal)** |
| **True Negatives (TN)** | **330.4 (mismo que normal)** |
| **False Negatives (FN)** | **31.4 (–23.0)** |
| **Accuracy** | **0.907 (+0.039)** |
| **Precision** | **0.897 (+0.011)** |
| **Recall** | **0.872 (+0.094)** |
| **F1** | **0.884 (+0.056)** |
| **AUC** | **0.967 (+0.037)** |
| **Average Precision (AP)** | **0.954 (+0.040)** |

**El modelo de random forest es claramente superior en todas las métricas al modelo de regresión logística normal, manteniendo la tasa baja en falsos positivos, pero recuperando muchos más verdaderos positivos, logrando además la AUC más alta (0.967)**

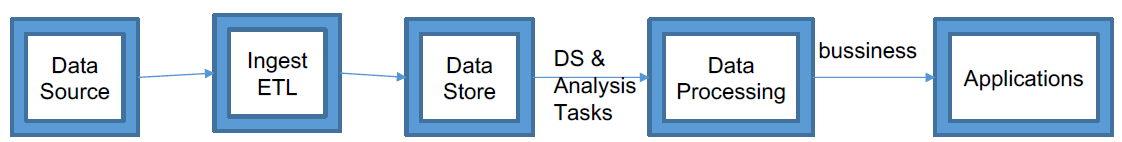
**Análisis y conclusiones**

1. **Regularización en regresión logística**
   1. Cuando las variables (p. ej. hipertensión, diabetes) son demasiado predictivas, el modelo logístico tiende a asignarles coeficientes muy grandes, lo que puede dar lugar a sobreajuste y pobre generalización en datos nuevos.
   2. La regularización (penalización de la normal del vector de coeficientes) reduce estos coeficientes, forzando al modelo a repartir parte del peso predictivo entre las demás variables (PC1, PC2, PC3, sexo, actividad física)
   3. Esto puede traducirse en mayor robustez frente a cambios leves en la distribución de la población y evita que un solo factor de riesgo acapare toda la decisión.
2. **Comparación práctica**
   1. Si la prioridad es maximizar la recuperación de casos positivos (recall), sin importar un aumento en falsos positivos, la regresión regularizada es una buena alternativa en comparación a la normal
   2. Si se busca balancear errores y se dispone de suficiente potencia computacional, **Random Forest** ofrece un rendimiento marcadamente superior (mejores AUC, F1 y AP), identificando más casos de riesgo cardiovascular con muy pocos falsos positivos.
3. **Elección final**
   1. **Entorno clínico o screening**: si se quiere identificar la mayor cantidad posible de pacientes en riesgo (aceptando más falsos positivos), optar por Random Forest o regresión regularizada.
   2. **Entorno de diagnóstico confirmatorio**: donde el costo de un falso positivo es alto (p. ej. pruebas invasivas), la regresión logística “normal” o Random Forest ajustado para baja FP podría ser preferible.

**Tecnología**

**Desarrollo del proyecto**

Debido a la naturaleza del proyecto, este se desarrolló siguiendo una arquitectura batch, utilizando tanto el servicio de almacenamiento de AWS S3 como el ambiente local de cada integrante del equipo, es decir, el computador personal de cada uno, debido a que el volumen de datos no es tan grande como para requerir una infraestructura tecnológica más robusta como el resto de servicios en la nube ofrecidos por Amazon para el resto de las etapas de desarrollo.



La fuente de datos y la naturaleza de los mismos es de tipo batch, ya que provienen de un archivo .csv descrito anteriormente en la sección *descripción de los datos*. Dado que dicho archivo contiene una tabla con un esquema bien definido para sus columnas, se trata de datos estructurados, provenientes de una consulta a una bodega de datos (o *Data WareHouse*) tradicional.

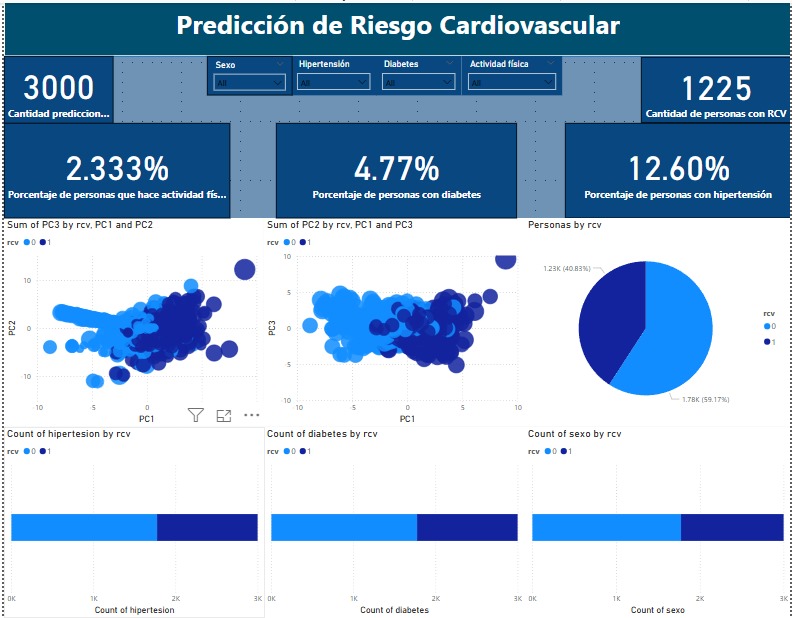
La ingesta de los datos se hizo de forma manual a AWS S3 a la zona *raw* de un bucket creado especialmente para este proyecto, por lo que dicha ingesta es de tipo batch.

Así mismo, los datos son almacenados en una lago de datos (o *datalake*) en el servicio de AWS S3 donde, en el mismo bucket mencionado anteriormente, también se tiene una zona *trusted* para los datos que ya fueron preparados para su procesamiento y una zona *refined* donde se almacenan los datos ya procesados, es decir, que son el resultado entregado por el modelo que se desarrolló; y también está almacenado el modelo ya entrenado para que ambas cosas (resultados y modelo) estén listas para su uso, ya sea para realizar nuevos procesamientos o para la visualización de los datos resultantes.

Hablando de la preparación y procesamiento de los datos, estás dos fases del desarrollo se hicieron en un entorno local de python, utilizando librerías como pandas y numpy para el uso y manipulación de los datos en dataframes y matrices respectivamente; y sklearn para la creación, entrenamiento y evaluación de los modelos de machine learning, de los cuales se eligió el modelo de *random forest* como el mejor, tras evaluar y comparar cada uno de los modelos como se mencionó anteriormente, y es el que será expuesto en la zona *refined* junto a sus resultados.

**Despliegue del proyecto**

Finalmente, la visualización de los datos se hace mediante un archivo de Power BI, el cual toma los datos resultantes del modelo definitivo (el desarrollado con *random forest*) de un archivo .csv descargado desde AWS S3 y alojado en la zona *refined* para tener una visualización descriptiva de tipo BI de los resultados arrojados por el modelo desarrollado.



El archivo de Power BI podría conectarse directamente a AWS S3 para cargar los datos desde la zona *refined*, pero, debido a las limitaciones de la licencia de AWS Academy con la que se cuenta, no se puede mantener esta alternativa por mucho tiempo debido a que una de las credenciales utilizadas para realizar dicha conexión (el token secreto) tiene una duración limitada ya que se trata de un token temporal.

Ante esta limitación, surge la pregunta de cómo sería el despliegue de este proyecto en un entorno de producción real. Pues bien, como se explicó con anterioridad, se tendría que contar con una licencia estándar o empresarial de AWS para evitar las limitaciones de una licencia educativa y ya con esto se podría crear un ODBS que se conecte al bucket en AWS S3 y, más específicamente, a la zona *refined* para que, desde el archivo de Power BI, se puedan cargar los datos y mostrar el tablero siempre que los usuarios finales lo necesiten. En ese sentido, se podría crear una aplicación web con un tablero de Power BI embebido que permita consultar dicho tablero a cualquier hora en cualquier momento (es decir, una disponibilidad 24/7) por cualquier persona que cuente con las credenciales de acceso.

**Conclusiones**

* El EDA permitió identificar variables redundantes, seleccionar atributos clave para el modelo y establecer relaciones estadísticamente significativas entre factores clínicos y el riesgo cardiovascular.
* El análisis PCA permitió simplificar la información clínica conservando su estructura esencial y facilitando la identificación de patrones. Las tres componentes principales extraídas resumen factores clave del riesgo metabólico, la función renal y la presión arterial, y pueden utilizarse como indicadores clínicos de alerta temprana y como variables resumen en modelos predictivos más robustos y explicables.
* El modelo de Random Forest superó a los modelos de regresión logística en todas las métricas evaluadas, destacándose por su mayor capacidad para identificar correctamente casos positivos sin aumentar la tasa de falsos positivos, mostrando ser una herramienta eficaz para escenarios donde la sensibilidad es importante.
* La inclusión de regularización en la regresión logística ayudó a mitigar posibles problemas de sobreajuste, distribuyendo mejor la influencia de las variables predictoras, aunque con una leve reducción de desempeño global.
* La combinación de variables clínicas (hipertensión, diabetes, sexo, actividad física) con componentes principales extraídas del análisis de datos mejora la capacidad predictiva de los modelos para riesgo cardiovascular.
* La visualización ayudó a validar estos patrones y resaltar la necesidad de un modelo predictivo robusto y clínicamente interpretable. Un ejemplo de esto es que, gracias al tablero, es posible detectar que, de las 3000 predicciones de RCV realizadas por el modelo, hubo 1225 personas con RCV en los próximos 5 años. Además, de esas personas que se evaluaron sólo el 2.33% realizan actividad física, 4.77% sufren de diabetes y 12.60% sufren de hipertensión. Así mismo, algo interesante que se puede notar es que toda persona con dichas aflicciones inmediatamente el modelo las clasifica con RCV, estando acorde con la realidad.

**Referencias**

[1] Mendias Benítez, C., Chain de la Bastida, J. L., & Ponce Martín, R. (s.f.). *Riesgo cardiovascular (RCV) de una población de pacientes en hemodiálisis. Comparación de los estudios SCORE, REGICOR y DORICA*. Revista SEDEN, (50), 27–35.<https://www.revistaseden.org/files/3453_Articulo%2050.pdf>

[2] NUEVA EPS. (s.f.). Programas de Promoción y Prevención. NUEVA EPS. <https://www.nuevaeps.com.co/programas-promocion-prevencion>

[3] Jaadi, Z. (n.d.). Principal Component Analysis (PCA): A Step-by-Step Explanation. Built In. <https://builtin.com/data-science/step-step-explanation-principal-component-analysis>

[4] Amazon Web Services. (s.f.). ¿Qué es la regresión logística?. AWS. <https://aws.amazon.com/es/what-is/logistic-regression/>

[5] IBM. (s.f.). *¿Qué es el bosque aleatorio?*. IBM Think.<https://www.ibm.com/mx-es/think/topics/random-forest>

[6] QuestionPro. (s.f.). Ciclo de vida de los datos: Qué es y qué etapas tiene. <https://www.questionpro.com/blog/es/ciclo-de-vida-de-los-datos/>

[7] IBM. CRISP-DM help overview. IBM Documentation. Disponible en:<https://www.ibm.com/docs/es/spss-modeler/saas?topic=dm-crisp-help-overview>