MACHINE LEARNING APLICADO A LA ESTIMACIÓN DE EVENTOS AGUDOS EN POBLACIÓN EN CONDICIÓN DE RIESGO CARDIOVASCULAR -RCV

Christian Felipe Alzate Cardona

Monografía presentada para optar al título de Especialista en Analítica y Ciencia de Datos

Tutor
Natalia Romero Rios, Magíster (MSc) en Estadística

Universidad de Antioquia
Facultad de Ingeniería

Especialización en Analítica y Ciencia de Datos

Medellín, Antioquia, Colombia

2021

Cita	(Alzate, 2021)
Referencia **Estilo APA 7 (2020)**	Alzate, CF., (2021). Machine Learning aplicado a la estimación de riesgo ante eventos agudos en población en condición de riesgo cardiovascular - RCV [Trabajo de grado especialización]. Universidad de Antioquia, Medellín, Colombia.

Especialización en Analítica y Ciencia de Datos, Cohorte II.

Universidad de Antioquia - www.udea.edu.co

Rector: Jhon Jairo Arboleda Céspedes

Decano/Director: Jesús Francisco Vargas Bonilla

Jefe departamento: Diego José Luis Botía Valderrama

TABLA DE CONTENIDOS

1. RESUMEN EJECUTIVO

La motivación es buena parte influenciada por el contexto de pandemia que se vive desde el año 2020, a causa de la expansión a nivel mundial del virus COVID, donde los servicios de salud se concentraron en atender los pacientes infectados, generando una desantención al resto de enfermedades incluidas las no trasmisibles, como es el caso de las enfermedades cardiovascualres y los eventos agudos relacionados con estos.

Debido a esta capacidad reducida en la atención causada volcamiento de los servicios hacia el fenómeno COVID, se vuelve cada vez más valiosa una estratificación del riesgo en el resto enfermedades críticas y en especial una priorización en la atención de los pacientes más propensos a complicaciones en ellas, para un uso adecuado de la capacidad restante.

Nos encontramos entonces con la oportunidad y la necesidad de aprovechar los datos que se captan en una compañía prestadora de servicios de salud, de forma que puedan contribuir a la anticipación de las decisiones que se toman con base en la información del pasado y entregar nuevas soluciones analíticas de valor que contribuyan al uso pertinente de los recursos disponibles a partir del conocimiento del negocio, sus procesos y necesidades.

Para el desarrollo se cuenta con un grupo anónimo de pacientes con alguna de las tres patologías siguientes, diabetes mellitus, hipertensión arterial o enfermedad renal crónica en estadio 1 al 4, considerándose población en condición de riesgo cardiovascular (RCV); a estos se les indagó acerca de su estado clínico al momento de presentar o no una hospitalización o una urgencia, cruzando con un conjunto de variables que identifican la posibilidad o no de tener comorbilidades adicionales, utilización de servicios y variables de hábitos de vida.

Se estimaron diferentes modelos de clasificación, se decidió por la regresión logística por su poder explicativo y sus métricas superiores al resto de los modelos. Obteniendo con este resultado una capacidad predictiva adecuada y que superó los requerimientos del negocio, con una sensibilidad y especificidad del 81% para la regresión con los mejores hiperparámetros y un cutoff optimo en un escenario de una clasificación dicotómica.
Además, como se decidió hacer una clasificación multicategórica de la probabilidad de la regresión también se cuenta con la bondad de tener un AUC del 80% generando los segmentos de dichas categorías mediante clustering, garantizando la separación más adecuada en función de la distribución de las probabilidades.

El modelo contenerizado, la documentación en detalle se encuentra en el siguiente repositorio:

https://github.com/christianfelipealzatecardona/RCV_Monografia.git

2. DESCRIPCIÓN DEL PROBLEMA

La priorización de pacientes es importante en la atención de eventos de salud, dado que el crecimiento de los mismos hace que se requiera la acomodación con el objetivo de salvar más vidas y utilizar los recursos del sistema de salud de forma eficiente. El sector salud ha utilizado técnicas de machine learning con el fin de apoyarse en esta labor, usando varias técnicas de predicción, detección, clasificación, de forma supervisadas y no supervisadas para el tratamiento de enfermedades, estudios clínicos y recomendación de patrones de atención de los mismos (O.H. Salman, 2021). Éstas técnicas fueron usadas también como apoyo para la atención a la pandemia por SARS-CoV-2, donde se reporta el uso de Deep learning para el desarrollo de medicamentos (Deepthi K., 2021), técnicas de clusterización para ayudar a enfocar los recursos (Reem, 2021) y de predicción para determinar efectos graves en los individuos (Lam, C. et al., 2021).

A raíz de la pandemia por Coronavirus, el crecimiento exponencial de los casos generó un foco en los pacientes de esta enfermedad, disminuyendo los recursos destinados a la atención de patologías diferentes, entre ellas las personas afectadas por enfermedades cardiovasculares, las cuales representan la primera causa de muerte a nivel mundial (Zhao et al., 2021). Diversos estudios han utilizado técnicas de Machine Learning como XGBoost models, elastic net, o Regresión logística para predecir mortalidad y, por ende, priorizar (Jing et al., 2020) (Agrawal et al., 2021). Se reporta que los determinantes más frecuentemente encontrados como relevantes, en estudios de desenlaces mortales de enfermedades cardiovasculares son género, raza, estado marital, ocupación, ingresos, circunferencia de cintura e índices hematológicos severos (Zhao et al., 2021), (Jing et al., 2020), (Agrawal et al., 2021).

En el presente trabajo se explora la utilización de algoritmos de machine learning como método que facilite la gestión anticipatoria del riesgo en salud de pacientes con riesgo cardiovascular en una compañía aseguradora. Por ello este acercamiento se hace principalmente relevante en las poblaciones más vulnerables, en este caso se aborda el riesgo de un paciente con deterioro cardiovascular de sufrir un evento agudo dentro de su patología.

2.1 PROBLEMA DE NEGOCIO

Las enfermedades no transmisibles -ENT-, que son definidas por la Organización Mundial de la Salud como patologías de larga duración que resultan de la combinación de factores genéticos, fisiológicos, ambientales y conductuales. Las ENT matan a más de 41 millones de personas al año, es decir, el 71% de las muertes que se producen en el mundo; de estas muertes, 15 millones son prematuras, pues se dan en población que se encuentra entre los 30 y los 69 años. El consumo de tabaco, la inactividad física, el uso nocivo del alcohol y las dietas malsanas aumentan el riesgo de morir de una ENT (Organización Mundial de la Salud, 2018).

La Enfermedad Cardiovascular, hace referencia a un conjunto de enfermedades (hipertensión arterial, cardiopatía coronaria, accidente cerebrovascular, enfermedad vascular periférica, insuficiencia cardíaca, cardiopatía reumática, cardiopatía congénita y miocardiopatías) que afectan el corazón, los vasos sanguíneos y el resto del organismo, principalmente el cerebro, los riñones y los miembros inferiores. Entre los eventos agudos que pueden desencadenar estas enfermedades están: el infarto de miocardio y el accidente cerebrovascular (trombosis, embolia y hemorragia cerebral). Son muy graves y la principal causa de muerte sobre todo en países desarrollados.

Los pacientes con mayor propensión a las enfermedades anteriormente mencionadas y en consecuencia a sus complicaciones son los que medicamente son categorizados en Riesgo Cardiovascular (RCV). Estos pacientes son aquellos que desde sus condiciones de salud tienen el perfil de riesgo para sufrir una de estas enfermedades, los factores que influyen según el criterio médico en dicho riesgo y que están consignados en los sistemas de información de la compañía aseguradora son: edad, sexo, raza, antecedentes familiares, hipertensión, colesterol, diabetes, tabaquismo, sedentarismo, alcohol, ansiedad, estrés. Estas fueron entonces las variables descriptoras de partida que se usaron para la predicción del fenómeno (Evento cardiovascular agudo).

De los usuarios atendidas por las prestadoras de servicios, el 13.5% tienen riesgo cardiovascular (RCV), donde la mayoría presentan hipertensión y se encuentran en adultez (29 a 59 años) y vejez (mayor o igual de 60 años) como se observa en Figura 1 Cantidad de usuarios con RCV.

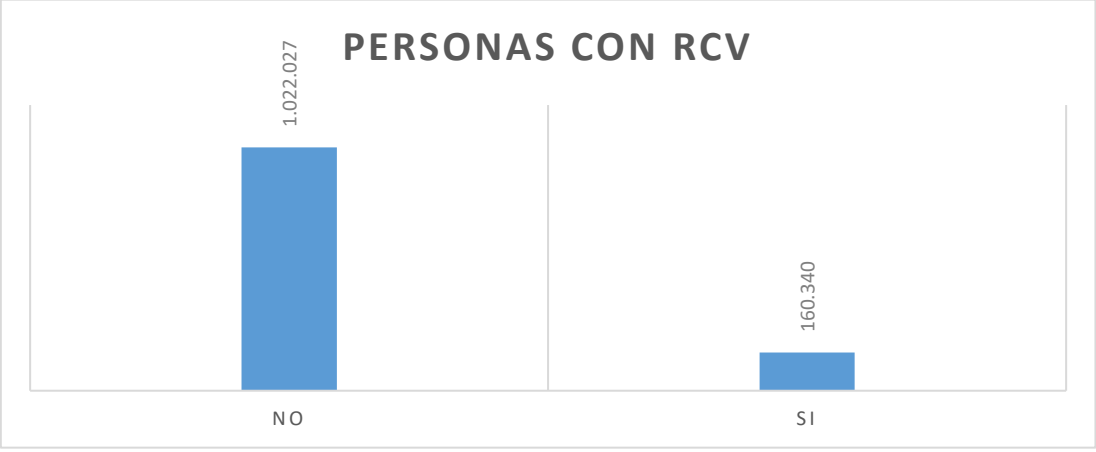


Figura 1 Cantidad de usuarios con RCV

Fuente: Elaboración propia con datos de la compañía aseguradora

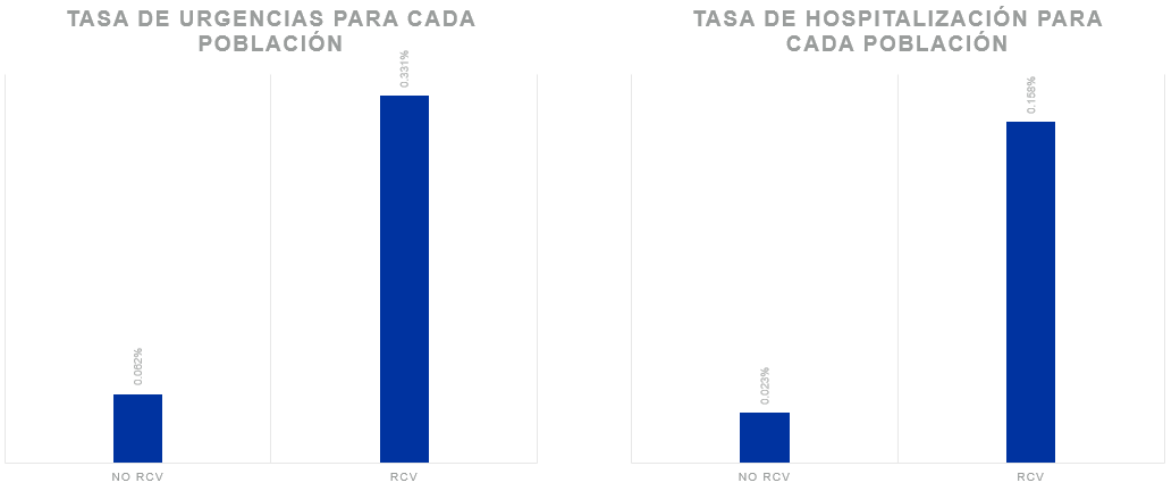


Figura 2 Tasas de eventos agudos según población

Fuente: Elaboración propia con datos de la compañía aseguradora

Como se evidencia en la Figura 2 Tasas de eventos agudos según población, si bien las personas con RCV son minoría, tienen una utilización de eventos de agravamiento mucho mayor que el resto. Es por esto se hace relevante construir un modelo que permita priorizar a las personas y que desencadene en la formulación de estrategias que permitan evitar hospitalizaciones y urgencias en los mismos.

2.2 APROXIMACIÓN DESDE LA ANALÍTICA DE DATOS

Como se ilustró en la sección anterior, el poder diseñar estrategias que permitan atender a la población basados en el nivel de riesgo, puede generar beneficios en los pacientes y en las instituciones prestadoras de salud. Sin embargo, la estimación del nivel de riesgo de un paciente involucra varias variables que para una persona o reglas se hace difícil o imposible de estimar. Por tal motivo, se propone una aproximación con técnicas analíticas que identifiquen la interacción de todas las variables que estén disponibles y sean pertinentes para mejorar la gestión el riesgo en salud.

Para esto se realiza una clasificación binaria para predecir con base a un conjunto de variables clínicas si un paciente en condición de riesgo cardiovascular es susceptible de presentar un evento agudo (una hospitalización o una urgencia).

Para la implementación de este modelo, se siguió la metodología iterativa Crisp-DM, la cual se ilustra en la Figura 3 Metodología Crisp-DM, siguiendo los pasos a continuación:

1. Entendimiento de negocio: Se realizó la selección y definición del problema con un panel de expertos interdisciplinarios, donde se establecieron los objetivos y alcances de este, además de evaluar la disponibilidad y calidad de la información
2. Entendimiento de los datos: Por medio de analítica descriptiva se encontraron las variables que además de tener relación de negocio, presentaran una relación numérica con las variables resultados
3. Preparación de los datos: Se realizaron los procesos de ETL para garantizar que los datos tuviesen la calidad y estructura requerida para la modelación
4. Modelado: Se realizó la evaluación de diferentes modelos y parámetros que permitiesen tener los resultados más acertados
5. Evaluación: Se evalúa si el modelo cumple con los parámetros de desempeño y los objetivos de negocio. En caso de no presentarlos, se realizan los ajustes pertinentes desde el paso 1
6. Puesta en producción: Se disponibiliza para que los usuarios puedan consumirlo

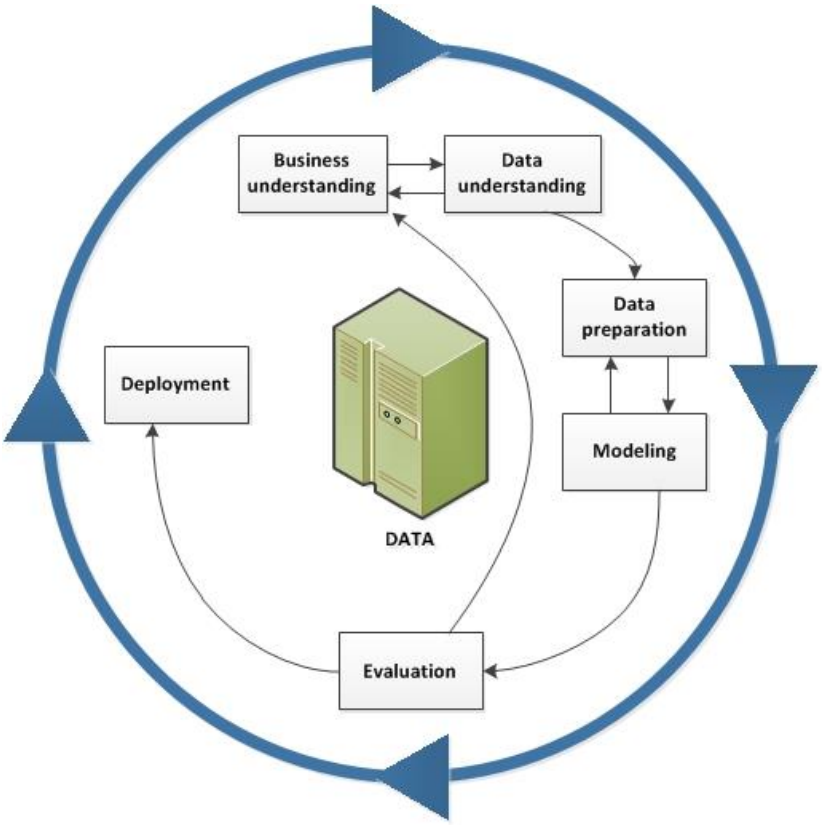


Figura 3 Metodología Crisp-DM

Tomado de (IBM, 2021)

2.3 ORIGEN DE LOS DATOS

La compañía asegurada dispone a las prestadoras de servicios de salud, sistemas transaccionales para el registro, actualización y almacenamiento de los datos, obtenidos al momento de la prestación del servicio; la integración de los distintos orígenes se materializa mediante procesos ETL (Extraer-Transformar-Cargar), transformando los datos de acuerdo con criterios de calidad y finalmente se cargan a una bodega de datos (*Data Warehouse*), regidos por políticas de gobierno y dominio de datos.

Los datos representan un grupo anónimo de pacientes, que están anotados según criterio médico con riesgo cardiovascular -RCV, además contiene variables sociodemográficas, comorbilidades, hábitos de vida y agrupamientos resultados de laboratorio y comportamentales como la utilización de servicios de salud como las urgencias y hospitalización; estos datos son obtenidos en las diferentes instituciones las prestadoras de servicios de salud.

La temporalidad de la selección del conjunto de datos estuvo entre 2021-05-01 y 2021-10-31, este periodo de tiempo ubica en un escenario postpandemia, donde la prestación de servicios y continuidad de los tratamientos necesarios de esta población se retomaron. Adicionalmente con el fin de disponer en un repositorio externo a la compañía, el periodo máximo autorizado para la extracción es de 6 meses dada las políticas internas.

2.4. MÉTRICAS DE DESEMPEÑO

Para evaluar este modelo, se proponen las siguientes métricas:

- **Métrica de negocio:**

La compañía no cuenta con un sistema de predicción de eventos agudos más allá de la tasa de hospitalización presentada en la Figura 2 Tasas de eventos agudos según población en la página 7, por lo que cualquier sistema que permita identificar poblaciones con probabilidades diferentes al azar, será un beneficio para la asignación de recursos y tratamiento de dichos pacientes.

- **Métrica de machine learning:**

Para estimar la capacidad de clasificación, se utilizará como base una matriz de confusión, que permitirá tener sus métricas asociadas: (Exactitud, Precisión, Sensibilidad, balance Accuracy Score y f1-score). Dado que la predicción de eventos agudos aún no tiene una aproximación dentro del negocio, este se basará en puntaje del área bajo la curva ROC, que se alinea a lo descrito en la bibliografía para este tipo de problemas (O.H. Salman, 2021).

3. DATOS

3.1 DATOS ORIGINALES

El conjunto de datos dispuesto contiene 51 columnas y 590.691 filas, cada una representando un usuario, la variable respuesta de presentar un evento agudo (hospitalización o urgencia), en un archivo en formato .csv y un peso de 128 Mb aproximadamente. Este dataset cuenta con registros anonimizados.

Tabla 1 Lista de campos Datos originales

Nombre de la variable	tipo	Descripción de la variable
MES_ID	int64	Consecutivo del mes en que se calcula los indicadores de la población
TIPO_SUCURSAL	object	Tipo de Sucursal (adscrita, propia)
REGIONAL	object	Descripción de la regional donde reside el afiliado (Medellín, Bogotá, entre otras)
SEXO_CD	object	Descripción del Sexo del afiliado (F= Femenino, M= Masculino)
EDAD	int64	Edad actual al momento del mes que se calculó el indica dor
GRUPO_ETARIO_DESC	object	Agrupación de la Edad en años (menor a 1, entre 1 y 4, 5 a 14, 15 a 18, 19 a 44, 45 a 49, 50 a 54, 60 a 64, 65 a 69, 70 a 74 y mayor a 75)
RAZA_DESC	object	Descripción de la raza (Mestizo, Blanco, Afroamericano, mulato, zambo, indígena)
IND_OBESIDAD	int64	Indicador Obesidad (1=Si o 0=No)
IMC	float64	Valor Índice de Masa Corporal y su categorización (sobrepeso, obesidad, normal y bajo peso)
COLESTEROL_TOTAL	object	Normal, limite alto y Alto
COLESTEROL_LDL_TXT	object	Normal, limite alto y Alto
COLESTEROL_HDL_TXT	object	Normal y Alto
TRIGLICERIDOS_TXT	object	Normal, limite alto y Alto
ESTADO_CIVIL_DESC	object	Estado Civil (Casado, Soltero, divorciado, Unión Libre, Viudo, Separado)
CODIGO_NIVEL_INGRESO_OP	object	Nivel de ingreso
CLASIFICACION_VALOR_INDICE_TXT	object	Descripción valor del índice de salud en más de 4 categorías (muy bajo, bajo, medio, alto, muy alto y sin información)
CLASIFICACION_POBLACION_DESC	object	Descripción valor del índice de salud (Sano Presuntivo, Enfermo Presuntivo, Condición de riesgo, Enfermo, SIN INFORMACION)
IND_CIGARRILLO	int64	Indica si el Afiliado Fuma (1=Si o 0=No)
IND_EJERCICIO	int64	Indica si el Afiliado hace ejercicio (1=Si o 0=No)
IND_LICOR	int64	Indica si el Afiliado toma alcohol (1=Si o 0=No)
IND_POSHOSPITALIZADO	int64	Indica si el Afiliado consultó después de una hospitalizado (1=Si o 0=No)

Nombre de la variable	tipo	Descripción de la variable
IND_POSURGENCIAS	int64	Indica si el Afiliado consultó a urgencias más de 3 veces en un periodo de tiempo corto (1=Si o 0=No)
IND_HIPERCONSULTANTE	int64	Indica si el Afiliado consultó por consume médico general más de 4 veces en un 30 día (1=Si o 0=No)
IND_ANTICOAGULANTE_NO_WARFA	int64	Indica si el Afiliado es anticoagulado por un medicamento diferente a la Warfarina (1=Si o 0=No)
IND_PROTECCIONRENAL	int64	Indica si el Afiliado tiene una enfermedad Renal Crónica estadio 1 a 4 (1=Si o 0=No)
IND_DIABETES	int64	Indica si el Afiliado tiene Diabetes Mellitus (1=Si o 0=No)
IND_HIPERTENSION	int64	Indica si el Afiliado tiene Hipertensión Arterial (1=Si o 0=No)
IND_EHC	int64	Indica si el Afiliado tiene Enfermedad hereditaria de la coagulación (1=Si o 0=No)
IND_GESTANTES	int64	Indica si el Afiliado está en embarazo (1=Si o 0=No)
IND_ASMA	int64	Indica si el Afiliado tiene Asma (1=Si o 0=No)
IND_DISLIPIDEMIA	int64	Indica si el Afiliado tiene Dislipidemia (1=Si o 0=No)
IND_VIH	int64	Indica si el Afiliado tiene VIH (1=Si o 0=No)
IND_EPOC	int64	Indica si el Afiliado tiene EPOC Mellitus (1=Si o 0=No)
IND_AUTOINMUNES	int64	Indica si el Afiliado tiene Autoinmunes (1=Si o 0=No)
IND_CANCER	int64	Indica si el Afiliado tiene Cáncer (1=Si o 0=No)
IND_INSUFICIENCIACARDIACA	int64	Indica si el Afiliado tiene una patología relaciona con Insuficiencia Cardiaca (1=Si o 0=No)
IND_CARDIOVASCULAR	int64	Indica si el Afiliado tiene enfermedad Cardiovascular (1=Si o 0=No)
IND_CEREBROVASCULAR	int64	Indica si el Afiliado tiene una patología Cerebrovascular (1=Si o 0=No)
IND_FOXIGENO	int64	Indica si el Afiliado es oxigeno requiriento (1=Si o 0=No)
IND_TUBERCULOSIS	int64	Indica si el Afiliado tiene Tuberculosis (1=Si o 0=No)
IND_HEPATITISC	int64	Indica si el Afiliado tiene Hepatitis C (1=Si o 0=No)
IND_PATO_MAMARIA	int64	Indica si el Afiliado tiene una patología Mamaria (1=Si o 0=No)
IND_ARTRITIS	int64	Indica si el Afiliado tiene Artritis (1=Si o 0=No)
IND_SIFILIS	int64	Indica si el Afiliado tiene Sífilis (1=Si o 0=No)
IND_EVE_DIF_RCV	int64	Indica si el Afiliado tiene una presentó un evento agudo diferente a los diagnósticos de RCV (1=Si o 0=No)
CANTIDAD_MARCA	int64	Número de comorbilidades anotadas tiene el afiliado
IND_FALLECIDO	int64	Indica si el Afiliado falleció (1=Si o 0=No)
CANT_HOSPITALIZACION	int64	Cantidad de Hospitalización en el último año
CANT_URG_ANIO	int64	Cantidad de Urgencias en el último año
AGRUPACION_COMPLICACION	object	Segmentación de Complicación (variable respuesta) por patologías
Y	int64	Indica si el Afiliado presentó un evento agudo, una hospitalización o urgencia (1=Si o 0=No)

Los datos estarán en el repositorio asignado por la universidad a cada estudiante y el uso de estos solo está permitido al autor con fines académicos y al docente del curso que para efectos de la calificación que debe comprobar la reproducibilidad del notebook.

3.2 DATASETS

El dataset obtenido por la compañía aseguradora llamado “Monografia.csv” se procede a cargar en un DataFrame de Pandas; el análisis exploratorio como primer contacto, lleva a comprender el problema a través de los datos, para esto se efectúan estadísticas descriptivas, correlaciones y gráficos. A demás permite verificar la calidad de estos y generar intuiciones con el fin de evaluar y seleccionar las características disponibles, si son relevantes o no para incluirlos en los datos de entrada. Finalmente se dividen aleatoriamente los datos en el 70% para entrenamiento de los modelos construidos y 30% restante se usa para la evaluación con datos desconocidos para ellos.

```
#Carga el Conjunto de Datos
df_original = pd.read_csv('/content/gdrive/My Drive/EACD-05- SEMINARIO/Monografia.csv',sep=',')
```

```
df_original.head()
```

RAZA_DESC	IND_OBESIDAD	IMC	IMC_TXT	COLESTEROL_TOTAL	COLESTEROL_LDL_TXT	COLESTEROL_HDL_TXT	TRIGLICERIDOS_TXT	ESTADO_CIVIL_DESC	C
BLANCO	0	30.62	Obesidad	Limite_Alto	Normal	Alto	Alto	Casado (a)	
SIN INFORMACION DESDE LA FUENTE	0	0.00	Sin Informacion	NaN	NaN	Alto	NaN	Casado (a)	
MESTIZO	0	27.47	Sobrepeso	Alto	Alto	Normal	Normal	Soltero (a)	
SIN INFORMACION DESDE LA FUENTE	0	31.22	Obesidad	Limite_Alto	Normal	Alto	Normal	Union Libre	
MESTIZO	0	18.73	Normal	Limite_Alto	Normal	Normal	Normal	Casado (a)	

Figura 4 Carga del archivo al Dataframe

Fuente: Elaboración propia

3.3 DESCRIPTIVA

Para este estudio la variable a predecir es el evento agudo, este indica si un paciente estuvo o no en los servicios hospitalización o urgencias por un diagnóstico relacionado a la patología; en la figura 2 se observa que la variable se encuentra altamente desbalanceada, es decir, son menos los pacientes (18785) que si presentaron este evento agudo, situación ideal para los pacientes y para la compañía, pero en los modelos de machine learning se sugiere aplicar estrategias para balancear y obtener mejores resultados al momento de entrenar los datos.

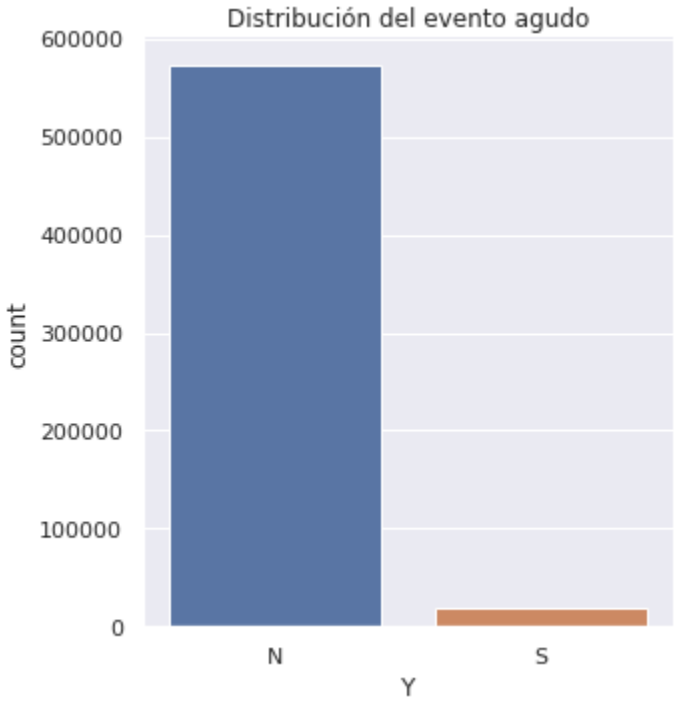
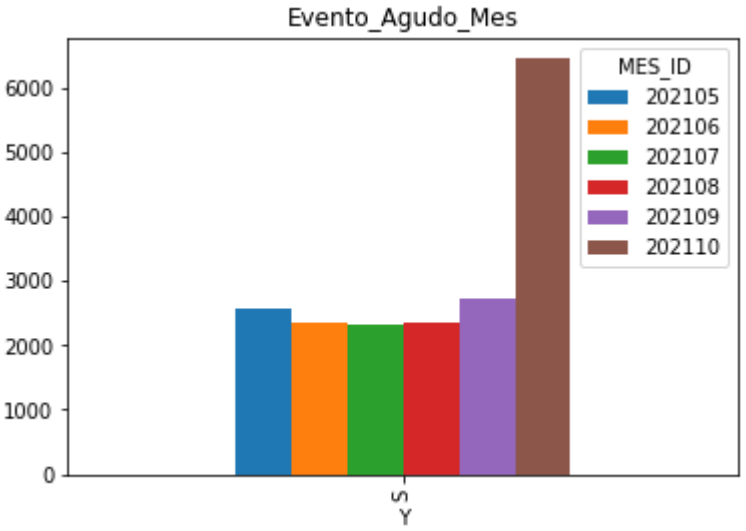


Figura 5 Variable a predecir (Evento Agudo)

Fuente: Elaboración propia

En el 2020 los servicios de salud se concentraron en atender la pandemia, además las personas en condición de RCV por el temor al contagio, disminuyeron la asistencia a controles incidiendo en que esta población no diera continuidad a los tratamientos médicos; por ende, se aumenta la presencia de eventos agudos (hospitalizaciones y urgencias por diagnósticos relacionados a RCV) ver Figura 6.



	MES_ID	Porcentaje
202105	2567	13.67
202106	2344	12.48
202107	2329	12.40
202108	2356	12.54
202109	2730	14.53
202110	6459	34.38

Según el programa de promoción y prevención del riesgo cardiovascular y estilos de vida saludable de la presidencia de la república, los hombres desde los 45 años y mujeres de 55 comienzan a ser más propensos a desarrollar una enfermedad cardiovascular; la edad y el sexo son factores de riesgo no modificables (Zhao et al., 2021) y se evidencia en la Figura 7 que la concentración para estas dos variables se encuentran entre los 40 y 80 años de vida para ambos géneros.

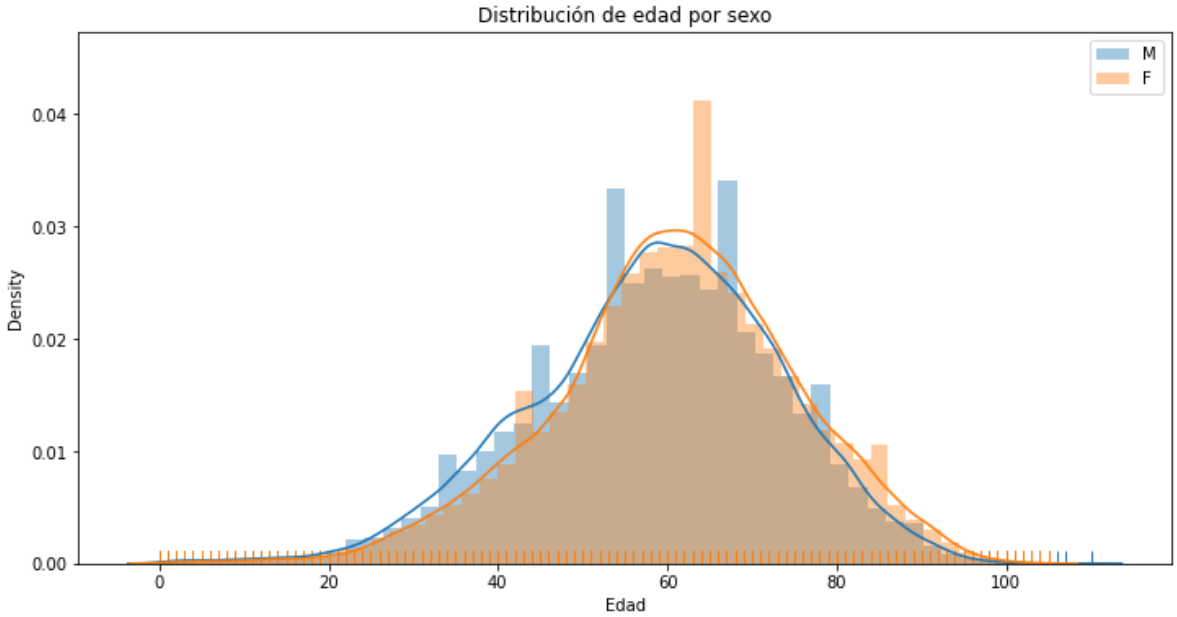


Figura 7 Distribución de edad y sexo

La actividad física es todo movimiento del cuerpo que hace trabajar a los músculos y requiere más energía que estar en reposo, como, por ejemplo: caminar, correr, bailar, nadar. El consumo de tabaco, la inactividad física, el uso nocivo del alcohol y las dietas malsanas aumentan el riesgo de morir de una enfermedad cardiovascular (Organización Mundial de la Salud, 2018). Los estilos de vida saludable hacen referencia a un conjunto de comportamientos o actitudes cotidianas que realizan las personas, para mantener su cuerpo y mente de una manera adecuada. La Figura 8 muestra que son menos las personas que indicaron hacer ejercicio.

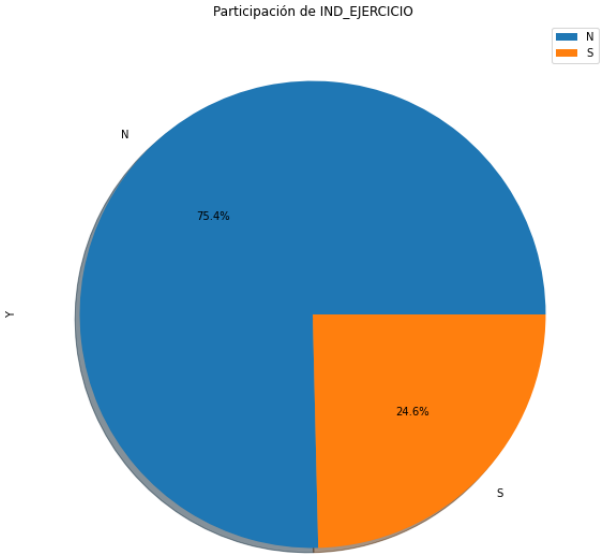


Figura 8 Distribución de personas que realizan ejercicio

Fuente: Elaboración propia

4. PROCESO DE ANALÍTICA

Después de la carga del dataset, comienza la etapa del preprocesamiento con el fin de comenzar el entendimiento de los datos.

4.1 PIPELINE PRINCIPAL

En la Figura 9 se encuentra ilustrado el flujo de trabajo general de los datos en el proyecto.

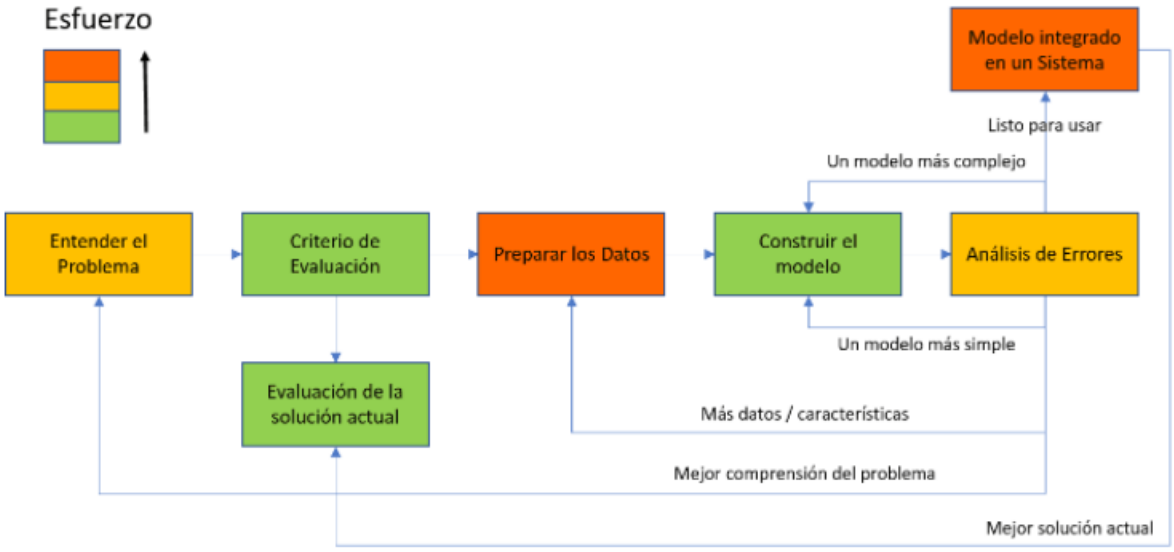


Figura 9 Proceso de Machine Learning

Tomado de: (IArtificial.net, 2020)

Descripción del flujo de trabajo:

1. Entender el problema: cada proyecto es diferente, y no se puede formular de la misma forma para cada proyecto; es importante tener claridad en la pregunta de negocio, interactuar con el equipo de expertos con distintas preguntas acerca del problema; adicionalmente entender los datos mediante el análisis exploratorio permite reforzar el objetivo.

2. Definir un criterio de Evaluación: acordar un criterio de evaluación con el negocio, para evaluar el desempeño del modelo es vital antes de comenzar, en este caso la calidad será medida con el puntaje de la matriz de confusión y buscando optimizar el AUC.

3. Evaluación de la solución actual: la compañía no cuenta con una solución y el problema tiene una complejidad alta, por lo que una solución simple no aplicaría en este caso; construir una solución de machine learning de forma automática.

4. Preparar los datos: en esta fase se invierte mucho tiempo, pero indispensable para utilizar los datos y estos nos ayuden a responder el problema por medio de algoritmos de machine learning; algunas de las situaciones que se deben realizar en esta etapa de preprocesamiento es la transformación de la variables explicativas en actividades como tratamiento de datos perdidos,

depuración datos atípicos, convertir variables categóricas a numéricas, normalización de datos, eliminación de variables sin importancia, entre otros.

5. Construir el modelo: elegir los algoritmos que se ajuste al problema, separar los datos de entrenamiento en un 70% y el 30% restante para realizar la validación sobre la precisión de los modelos, para finalmente seleccionar el mejor según la métrica y explicación del problema.

6. Análisis de Errores: Es importante comprender que tenemos que hacer para mejorar los resultados de machine learning (usar un modelo más complejo o simple, o requerimos más datos o más características). En esta fase es importante asegurar que el modelo es capaz de producir buenos resultados con datos nuevos.

7. Modelo integrado en un sistema: para maximizar la utilidad del machine learning el modelo se debe integrar en un sistema transaccional.

4.2 PREPROCESAMIENTO

El procesamiento inicia con la importación desde Google Drive y carga de datos en un DataFrame de pandas, utilizando el método info (), se encuentra información como el tipo de datos de cada variable y permite evidenciar cuenta a simple vista cuales contienen datos no nulos, como se muestra en la Figura 10.

```
7 display(df_original.info())
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 590691 entries, 0 to 590690
Data columns (total 53 columns):
 #   Column                      Non-Null Count   Dtype
---  ------                      --------------   -----
 0   [illegible]                 [illegible]      [illegible]
 1   MES_ID                      590691 non-null  int64
 2   TIPO_SUCURSAL               590691 non-null  object
 3   REGIONAL                    590691 non-null  object
 4   SEXO_CD                     590691 non-null  object
 5   EDAD                        590691 non-null  int64
 6   GRUPO_ETARIO_DESC           590691 non-null  object
 7   RAZA_DESC                   590691 non-null  object
 8   IND_OBESIDAD                590691 non-null  int64
 9   IMC                         560244 non-null  float64
 10  IMC_TXT                     590691 non-null  object
 11  COLESTEROL_TOTAL            460679 non-null  object
 12  COLESTEROL_LDL_TXT          431711 non-null  object
 13  COLESTEROL_HDL_TXT          590691 non-null  object
 14  TRIGLICERIDOS_TXT           448406 non-null  object
 15  ESTADO_CIVIL_DESC           590691 non-null  object
 16  CODIGO_NIVEL_INGRESO_OP     590691 non-null  object
```

Figura 10 Tipos de Datos del conjunto de datos

Fuente: Elaboración propia

Con el objetivo de visualizar los datos mediante gráficas, se realiza el reemplazo los valores de las variables numéricas a categóricas como se muestra en la Figura 11.

```
df_original[['IND_OBESIDAD', 'IND_POSHOSPITALIZADO',
       'IND_POSURGENCIAS', 'IND_HIPERCONSULTANTE',
       'IND_ANTICOAGULANTE_NO_WARFA', 'IND_PROTECCIONRENAL', 'IND_DIABETES',
       'IND_HIPERTENSION', 'IND_EHC', 'IND_GESTANTES', 'IND_ASMA',
       'IND_DISLIPIDEMIA', 'IND_VIH', 'IND_EPOC', 'IND_AUTOINMUNES',
       'IND_CANCER', 'IND_INSUFICIENCIACARDIACA', 'IND_CARDIOVASCULAR',
       'IND_CEREBROVASCULAR', 'IND_FOXIGENO', 'IND_TUBERCULOSIS',
       'IND_HEPATITISC', 'IND_PATO_MAMARIA', 'IND_ARTRITIS', 'IND_SIFILIS',
       'IND_EVE_DIF_RCV','IND_FALLECIDO','Y']]=df_original[['IND_OBESIDAD', 'IND_POSHOSPITALIZADO',
       'IND_POSURGENCIAS', 'IND_HIPERCONSULTANTE',
       'IND_ANTICOAGULANTE_NO_WARFA', 'IND_PROTECCIONRENAL', 'IND_DIABETES',
       'IND_HIPERTENSION', 'IND_EHC', 'IND_GESTANTES', 'IND_ASMA',
       'IND_DISLIPIDEMIA', 'IND_VIH', 'IND_EPOC', 'IND_AUTOINMUNES',
       'IND_CANCER', 'IND_INSUFICIENCIACARDIACA', 'IND_CARDIOVASCULAR',
       'IND_CEREBROVASCULAR', 'IND_FOXIGENO', 'IND_TUBERCULOSIS',
       'IND_HEPATITISC', 'IND_PATO_MAMARIA', 'IND_ARTRITIS', 'IND_SIFILIS',
       'IND_EVE_DIF_RCV','IND_FALLECIDO','Y']].replace(1,'S')
```

Figura 11 Reemplazo de variables numéricas a categóricas

Fuente: Elaboración propia

Una vez conocida la información de los tipos de datos, se procede a verificar por medio de un análisis estadístico del DataFrame, con el método describe () con el fin de identificar si hay problemas matemáticos como datos atípicos o grandes desviaciones, como se muestra en la Figura 12.

```
df_original.describe()
```

	MES_ID	EDAD	IMC	CANTIDAD_MARCA	CANT_HOSPITALIZACION	CANT_URG_ANIO
count	590691.000000	590691.000000	560244.00000	590691.000000	590691.000000	590691.000000
mean	202107.782467	59.835135	28.39217	3.099362	0.486556	0.248260
std	1.823284	14.654571	5.67420	1.189451	1.234460	0.710644
min	202105.000000	0.000000	10.00000	0.000000	0.000000	0.000000
25%	202106.000000	51.000000	24.89000	2.000000	0.000000	0.000000
50%	202108.000000	60.000000	27.67000	3.000000	0.000000	0.000000
75%	202110.000000	70.000000	31.08000	4.000000	1.000000	0.000000
max	202110.000000	110.000000	140.00000	12.000000	41.000000	56.000000

Figura 12 Estadísticos de las variables numéricas

Fuente: Elaboración propia

En análisis estadístico de los datos tipo object, se evalúan un conjunto diferente de estadísticas, unique refiere al número de observaciones distintas en la columna, top es el dato más frecuente que se produce y freq es la cantidad de veces que aparece el objeto top en la columna, las columnas que aparecen con NaN significa que estas métricas estadísticas en particular no se pueden calcular para ese especifico tipo de datos de columna, como se muestra en la Figura 13.

```
1 df_original[].describe(include='all')
```

	MES_ID	TIPO_SUCURSAL	REGIONAL	SEXO_CD	EDAD	GRUPO_ETARIO_DESC	RAZA_DESC	IND_OBESIDAD	IMC	IMC_TXT	COLESTEROL_TOTAL	C
count	590691.000000	590691	590691	590691	590691.000000	590691	590691	590691	560244.00000	590691	460679	
unique	NaN	2	12	2	NaN	12	8	2	NaN	5	3	
top	NaN	ADSCRITA	Medellin	F	NaN	Mayor de 75	MESTIZO	N	NaN	Sobrepeso	Normal	
freq	NaN	572160	379580	331089	NaN	90671	265709	581328	NaN	236724	306690	
mean	202107.782467	NaN	NaN	NaN	59.835135	NaN	NaN	NaN	28.39217	NaN	NaN	
std	1.823284	NaN	NaN	NaN	14.654571	NaN	NaN	NaN	5.67420	NaN	NaN	
min	202105.000000	NaN	NaN	NaN	0.000000	NaN	NaN	NaN	10.00000	NaN	NaN	
25%	202106.000000	NaN	NaN	NaN	51.000000	NaN	NaN	NaN	24.89000	NaN	NaN	
50%	202108.000000	NaN	NaN	NaN	60.000000	NaN	NaN	NaN	27.67000	NaN	NaN	
75%	202110.000000	NaN	NaN	NaN	70.000000	NaN	NaN	NaN	31.08000	NaN	NaN	

Figura 13 Estadísticas de las variables categóricas

Fuente: Elaboración propia

El dataset contiene datos faltantes se reemplaza los valores (sin información o nulos) por estructura que pueda ser utilizada para procesamiento posterior (NaN), que puede observarse en la Figura 14.

```
##según conocimiento de los datos, cambiamos la Representacion de los NA's
df_original= df_original.replace('?',np.NaN)
df_original= df_original.replace('Sin Informacion',np.NaN)
df_original= df_original.replace('Sin informacion',np.NaN)
df_original= df_original.replace('SIN INFO',np.NaN)
df_original= df_original.replace('SIN INFORMACION',np.NaN)
df_original= df_original.replace('SIN INFORMACION DESDE LA FUENTE',np.NaN)
df_original= df_original.replace('Sin Información',np.NaN)
df_original= df_original.replace('Sin Información desde la fuente',np.NaN)
df_original= df_original.replace('NA',np.NaN)
```

Figura 14 reemplazo de etiquetas que identifican valores faltantes

Fuente: Elaboración propia

Con base en los resultados obtenidos en el análisis descriptivo se procede a tomar las siguientes decisiones del dataset:

1. **Eliminación de variables sin relevancia:** con un experto de negocio se realiza la depuración de estas variables, considerándose que no le aportarían información al modelo, reduciendo el número de características de 53 a 40 columnas.
2. **Imputación de valores nulos:** con el experto de negocio se imputan datos faltantes, con el fin de no eliminar registros.
3. **Valores atípicos:** se elimina los registros 5950 con datos atípicos de la variable edad, que dando con un dataframe de 584741.
4. **Correlación entre variables numéricas:** no hay correlación entre las variables, por ende, no se requiere exclusión de ninguna variable numérica. La correlación no implica causalidad,

es decir, el hecho de que dos variables estén correlacionadas no necesariamente significa que una de ellas sea la causa de la otra. Sperman nos indica cuales variables presentan una correlación positiva o negativa, y cuales correlaciones se le atribuyen al azar ya que su significación es mayor a 0,05.

5. **Colinealidad:** Una vez se verifica la calidad de los datos, se procede a codificar las variables categóricas a numéricas por medio de la función get_dummies, eliminando siempre un factor y aumentando así el número de columnas a 63. Con el fin de evitar problemas de multicolinealidad.
6. **Estandarización de los datos:** técnica de preprocesamiento aplicada para expresar los valores numéricos de las variables explicativas en una misma escala, la estandarización a partir de la distribución Gaussiana con media cero y varianza unitaria, para esto se utiliza la librería de scikit learn StandarScaler.

Una vez se termina el preprocesamiento el dataframe se persiste al mismo repositorio de Google con el nombre 'Monografía_limpio.csv', con el fin de llevar a cabo el ciclo MLOps entrenar, registro y despliegue del modelo como servicio de Azure Machine Learning.

4.3 MODELOS

Para abordar este problema de clasificación en la primer iteracion se eligen 3 algoritmos con sus parámetros por defectos, incluidos en el paquete de sklearn: *QuadraticDiscriminant, GaussianNB, LogisticRegression*

Análisis discriminante cuadrático - (QuadraticDiscriminant) Es un clasificador estadístico que tiene la ventaja de ser fácilmente computable, no tiene hiperparámetros para ajustar, que es inherentemente multiclase y ha probado trabajar bien en la práctica. (Scikit-learn.org, 2021)

Clasificador bayesiano ingenuo - (GaussianNB) Es un algoritmo de aprendizaje supervisado que se basa en la aplicación del teorema de bayes, asumiendo la independencia condicional dentro de cada par de atributos. Su estimación es rápida, y dado que no asume dependencia, se evita problema derivados por la dimensionalidad. (Scikit-learn.org., 2021)

Regresión logística - (LogisticRegression) Es un algoritmo de predicción de resultado de variables categóricas, es decir, se puede utilizar para clasificación en variables binarias. Su salida es la probabilidad de las diferentes salidas de una variable de respuesta, computada utilizando una función logística. (Scikit-learn.org., 2021)

En la segunda iteración se entrenaron los siguientes modelos con el objetivo de comparar su desempeño frente a la primera iteración.

Random Forest Un bosque aleatorio es un metaestimador que se ajusta a varios clasificadores de árboles de decisión en varias submuestras del conjunto de datos y usa promedios para mejorar la precisión predictiva y controlar el sobreajuste. El tamaño de la submuestra se controla con el max_samples parámetro si bootstrap=True(predeterminado); de lo contrario, se utiliza todo el conjunto de datos para construir cada árbol.

BaggingClassifier Un clasificador de ensacado es un metaestimador de conjunto que ajusta los clasificadores base cada uno en subconjuntos aleatorios del conjunto de datos original y luego

agrega sus predicciones individuales (ya sea por votación o promediando) para formar una predicción final. Un metaestimador de este tipo se puede utilizar típicamente como una forma de reducir la varianza de un estimador de caja negra (por ejemplo, un árbol de decisión), introduciendo la aleatorización en su procedimiento de construcción y luego haciendo un conjunto a partir de él. [5]

4.4 MÉTRICAS

A continuación, se muestra un análisis comparativo entre estas métricas de desempeño para los diferentes modelos entrenados.

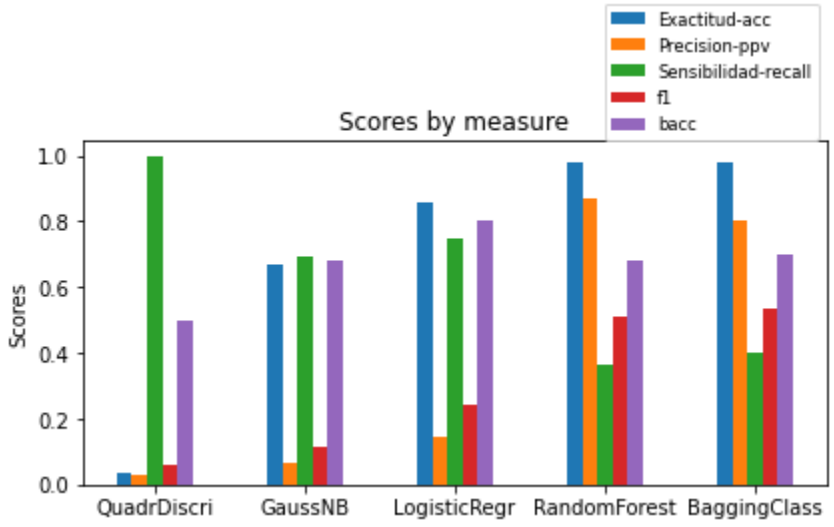


Figura 15 Resultados de las métricas de los modelos entrenados

Fuente: Elaboración propia

Al momento de aplicar algún algoritmo de clasificación supervisada y evaluar los datos, en especial si los estos no están balanceados, el modelo va a cometer dos tipos de errores, el tipo 1 falso positivo, ocurre cuando el modelo clasifica a un paciente que presenta un evento agudo cuando en realidad no lo presenta, y el error tipo 2 falsos negativos, donde el paciente realmente presenta el evento agudo pero el algoritmo predijo que no, que mejor para ello hacer el análisis mediante métricas que salen de la matriz de confusión.

- Accuracy (Exactitud): Cantidad de predicciones correctas; este no es siempre un buen indicador para validar que el algoritmo este haciendo una buena predicción.

- Precision: Porcentaje de casos positivos detectados.

- Recall (Sensibilidad): Proporción de casos positivos correctamente identificados.

- Especificidad: Casos negativos que se han clasificado correctamente.

- balance Accuracy Score y

- F1: ponderado entre la precisión y el recall.

Todas incluidas en la librería Metrics de Sklearn, se importan las siguientes métricas de Machine Learning, accuracy_score, balanced_accuracy_score, f1_score, precision_score y recall_score, y plot_confusion_matrix, classification_report y multilabel_confusion_matrix. Adicional la métrica de desempeño a evaluar los modelos fue AUC área bajo la curva ROC.

De las figuras 16 a la 20 registramos la matriz de confusión de cada uno de los modelos utilizados.

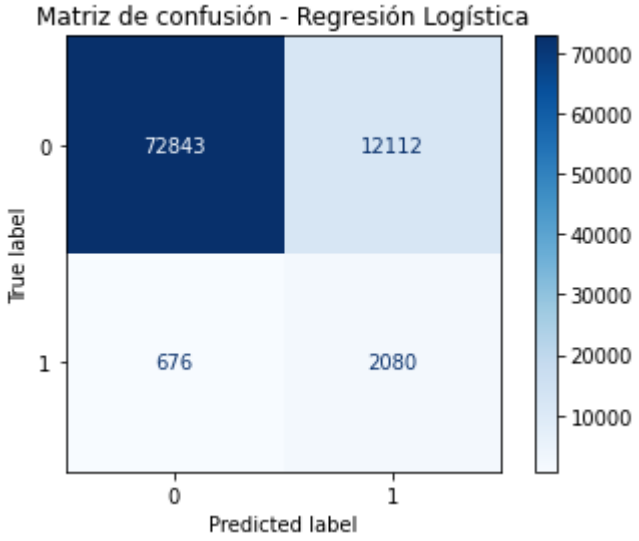


Figura 16 Matriz de confusión Regresión Logística

Fuente: Elaboración propia

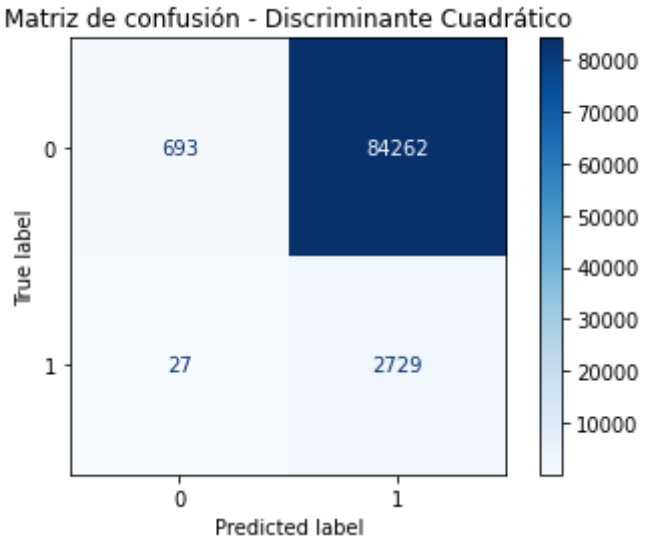


Figura 17 Matriz de confusión - Discriminante cuadrático

Fuente: Elaboración propia

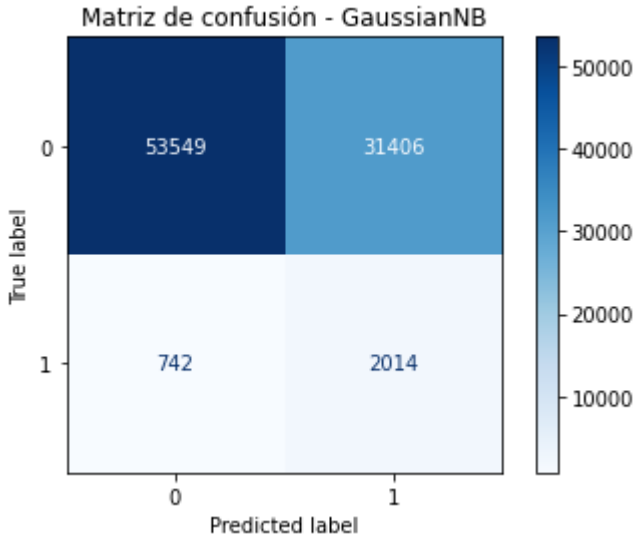


Figura 18 Matriz de confusión – GaussianNB

Fuente: Elaboración propia

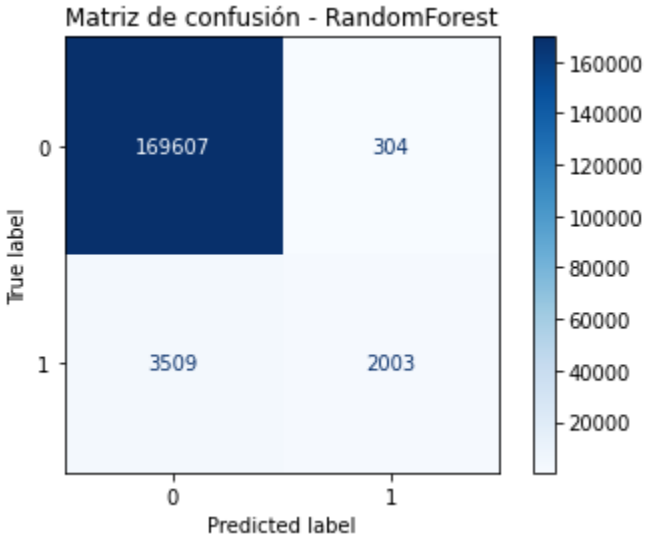


Figura 19 Matriz de confusión – RandomForest

Fuente: Elaboración propia

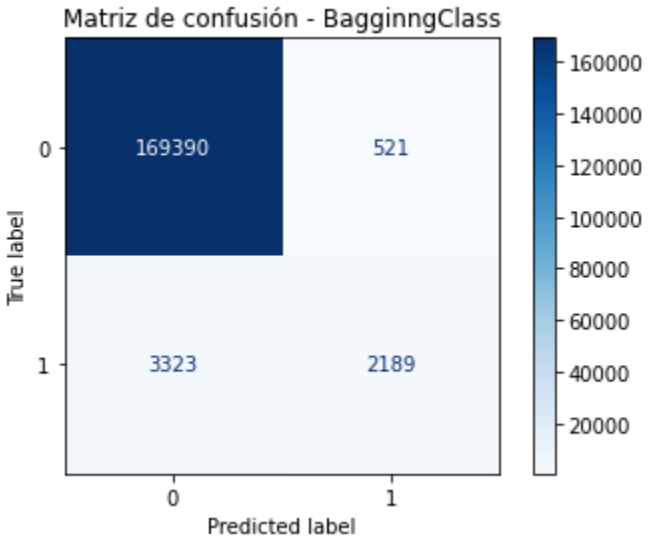


Figura 20 Matriz de confusión - BaggingClass

Fuente: Elaboración propia

5. METODOLOGÍA

5.1 BASELINE

Se escogen tres algoritmos de clasificación QuadraticDiscriminant, GaussianNB y LogisticRegression con sus hiperparámetros por defecto para tener un valor base (baseline), para comparar con los demás modelos.

5.2 VALIDACIÓN

Ya con la data preprocesado se realizó la estandarización y selección de los datasets de entrenamiento y prueba, utilizando la función traintestsplit de sklearn definiendo el 70% para entrenamiento y un 30% de validación, utilizando el método stratify teniendo en cuenta que la base está altamente desbalanceada. También durante el proceso de entrenamiento de los modelos se utilizó validación cruzada para lograr resultados más confiables.

La Figura 21 que muestra la curva ROC para el modelo de regresión logística, presenta un AUC del 80% que permite hablar de un buen poder predictivo del modelo ante diferentes cutoff y en consecuencia elegir un punto de corte que maximice la sensibilidad y la especificidad del clasificador y es superior a estudios referenciados dentro de metaanálisis (O.H. Salman, 2021).

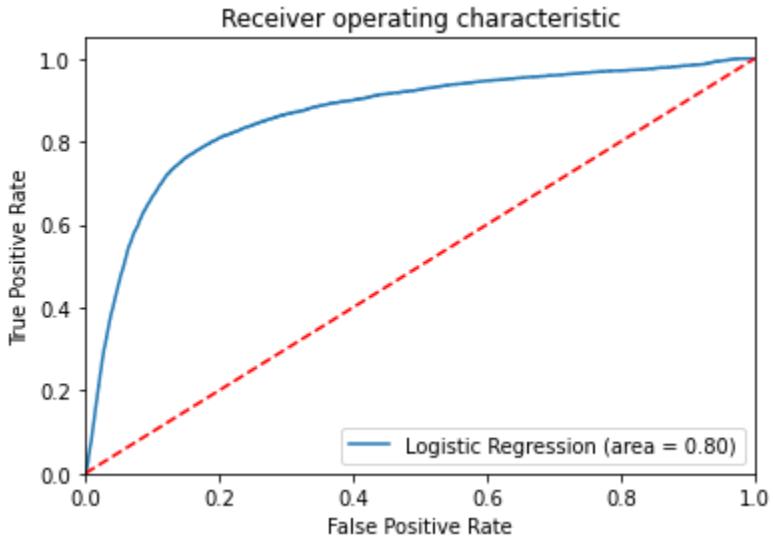


Figura 21 Curva ROC Regresión Logística

Fuente: Elaboración propia

Se decide recalcular el cutoff de la regresión logística con el objetivo de tener una sensibilidad y una especificidad balanceada, de tal manera que la precisión del modelo a la hora de predecir tanto a un individuo con riesgo de presentar un evento agudo como a uno que no lo presente sea igual, dicho cutoff se muestra en la Figura 22 Sensibilidad y especificidad Balanceada

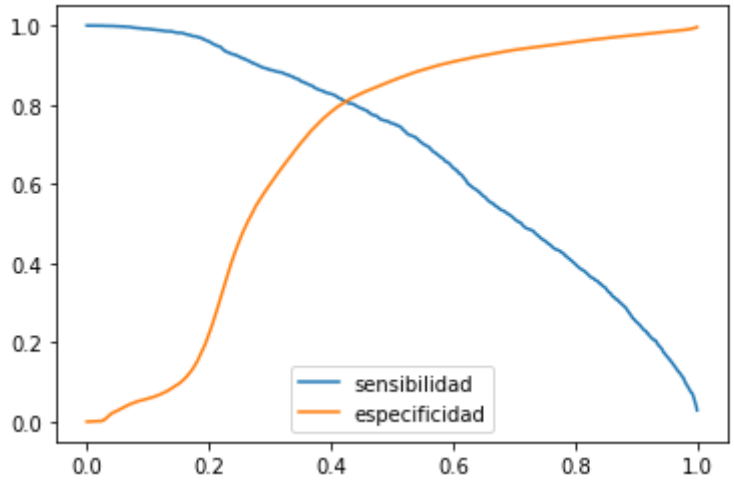


Figura 22 Sensibilidad y especificidad Balanceada

Fuente: Elaboración propia

5.3 ITERACIONES, EVOLUCIÓN y EVALUACIÓN CUALITATIVA

En el proceso iterativo posterior al primer acercamiento se tuvo como foco la estimación y comparación de varios modelos de clasificación en búsqueda del modelo con mejores bondades predictivas y al mismo tiempo con la capacidad de ser explicativo atendiendo a esta necesidad y en contraste con los otros modelos se eligió la regresión logística que también optimizada fue optimiza hiperparámetros por su posibilidad de proporcionar información extra sobre el fenómeno modelado, para esto se inspeccionó la significancia estadística de las regresoras, su magnitud y sentido como una aproximación a cuales de ellas tienen un mayor efecto inductor o protector en el desenlace.

Además de la etapa de clasificación en iteración final se agregó al modelo una capa de etiquetado que permitiera un detalle mayor al que podría dar una clasificación dicotómica, sin la cardinalidad del valor puro de la probabilidad y con el objetivo de servir como herramienta para la priorización en la operación. Para ello se efectuó un ejercicio de clustering donde a partir de los puntos medio entre los centroides generados por un modelo de KMeans y el cutoff óptimo anteriormente se logró separar en cinco niveles las zonas densas en el espacio de las probabilidades entregadas por la regresión logística aproximándonos con ello a grupos de personas con variables regresoras similares y del mismo modo grupos con un riesgo similar.

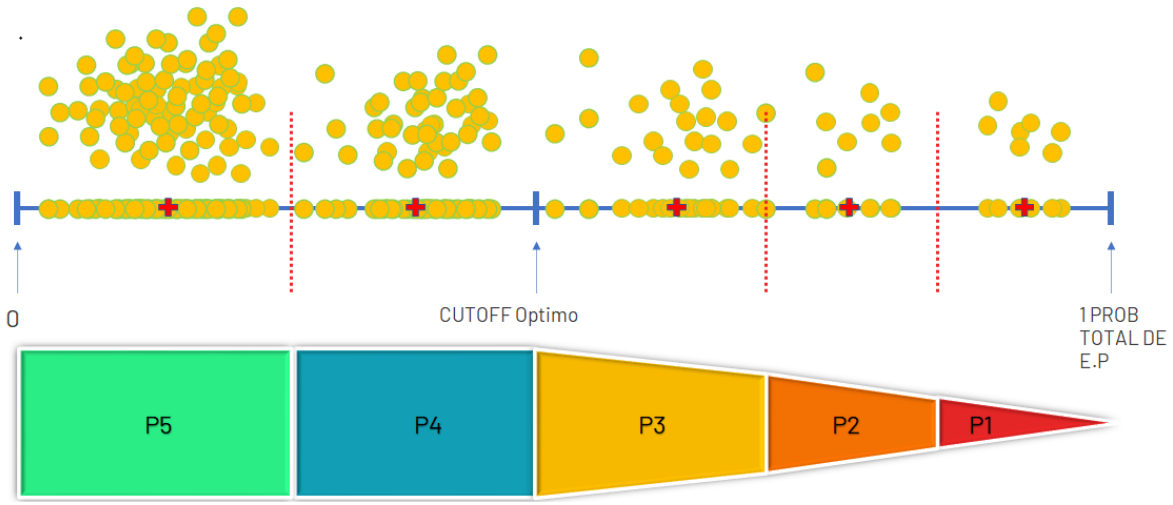


Figura 23 Creación de segmento de priorización por medio de Clustering

Fuente: Elaboración propia

5.4 HERRAMIENTAS

Para el desarrollo de este documento se utilizó:

1. Google Colab: herramienta principal utilizada para realizar los notebooks.

2. Librerías de Python:
    - Librería de datos (Pandas, Numpy)
    - Librería de Visualización (Matplotlib, Seaborn)
    - Librerías de Algoritmos (Scikit-learn)

3. Google drive: para almacenamiento y administración de la base de datos.

4. Visual Studio Code, utilizado para realizar la prueba del despliegue del modelo.

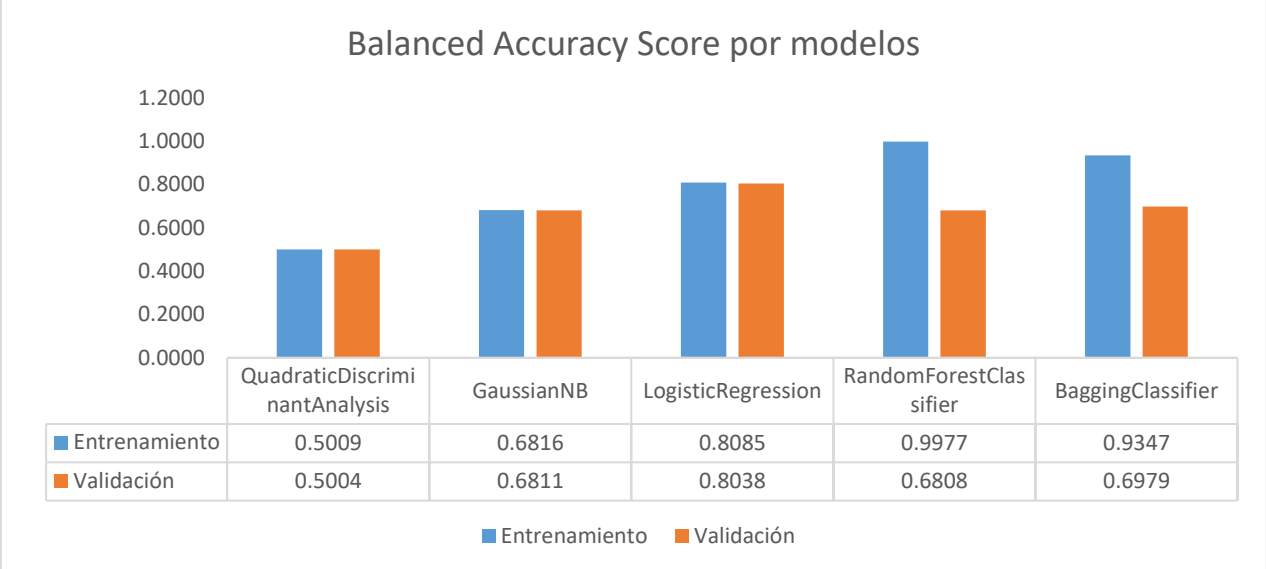
5. Servicios de AzureMl, utilizado para el ciclo de vida MLOps.

6. Github: repositorio para almacenar el proyecto.

6. RESULTADOS

6.1 MÉTRICAS

Se evaluaron los diferentes modelos usando el Balanced Accuracy Score, el cual se observa en la figura 24:



	QuadraticDiscriminantAnalysis	GaussianNB	LogisticRegression	RandomForestClassifier	BaggingClassifier
Entrenamiento	0.5009	0.6816	0.8085	0.9977	0.9347
Validación	0.5004	0.6811	0.8038	0.6808	0.6979

Figura 24 Balanced Accuracy Score por modelo

Fuente: Elaboración propia

Se elegio la regresión logística debido a que tanto en entrenamiento como en validación presenta un rendimiento bueno y consistente que nos habla de un modelo bien generalizado y sin problema de varianza.

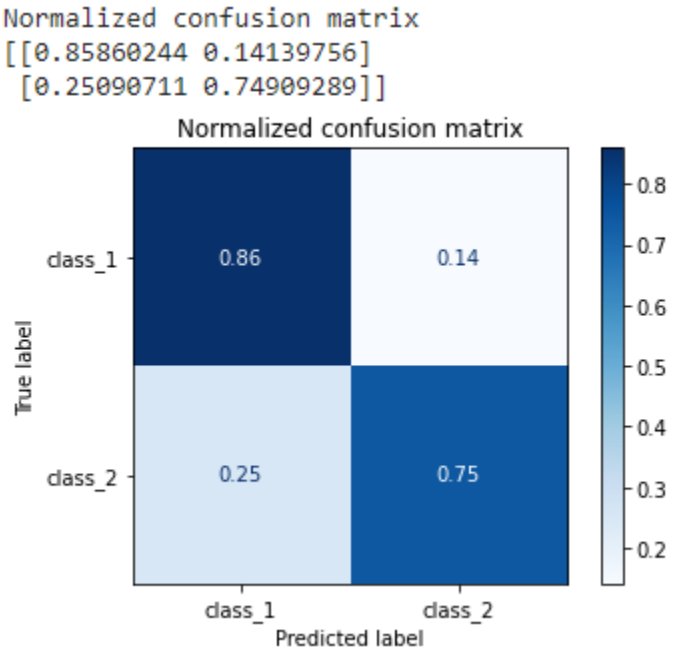


Figura 25 Sensibilidad y especificidad balanceada

Fuente: Elaboración propia

6.2 CONSIDERACIONES DE PRODUCCIÓN

Gracias a que la compañía cuenta con los servicios en la nube de Azure el modelo en cuestión podría ser fácilmente expuesto como una API que se integraría con los sistemas transaccionales en los que la priorización de los pacientes susceptibles a este fenómeno sea valiosa. Se vuelve indispensable generar un reentrenamiento periódico que para el caso podría hacerse mensualmente debido a que se trata de un fenómeno no tan cambiante. Además, se debe acompañar los reentrenamientos de un monitoreo constante de las métricas para garantizar la no degradación de la capacidad predictiva del modelo, también la inspección de la congruencia de dicho monitoreo con los criterios de los expertos clínicos y de negocio.

Con fines académicos en la cuenta del estudiante se despliega como servicio en la nube de Azure, el cual se ilustra en la captura de pantalla de la configuración, presente en la Figura 26. El modelo contenerizado, la documentación en detalle se encuentra en el siguiente repositorio:

https://github.com/christianfelipealzatecardona/DespliegueAzure.git

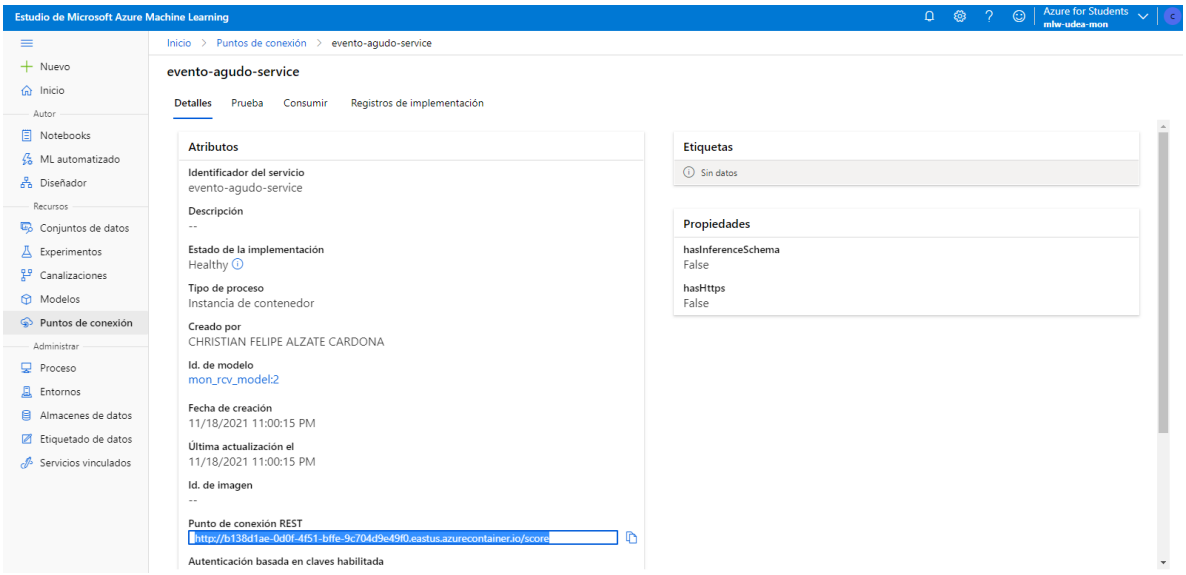


Figura 26 Configuración de Azure

7. CONCLUSIONES

En el presente trabajo se realizó un ejercicio de priorización de pacientes con Riesgo Cardiovascular, con enfoque en la utilización de servicios de urgencia y hospitalización, y utilizando diferentes metodologías de machine learning, de los cuales, por medio de las métricas de AUC y Accuracy, y poder explicativo, se eligió la regresión logística como mejor modelo para los datos.

A nivel de modelado la implementación de una regresión logística con la técnica stepwise con selección de variables basados en un criterio de información adecuada (AIC, BIC, etc) para obtener un modelo más parsimonioso y simple pero igual o más potente siendo precisos con las variables que mejor se exponen como casuales del desenlace, el evento agudo. Este modelo es adecuado comparando con metaestudios (Jing et al., 2020) (Agrawal et al., 2021) (O.H. Salman, 2021)

Dentro de los pasos a seguir luego de éste artículo es la implementación de este modelo, la cual permitirá hacer un uso pertinente de las capacidades en esta población riesgosa, pero en la mayoría de los casos fácilmente gestionable, que por el uso de este modelo, podrán aumentar en precisión. Esto requiere la construcción de herramientas complementarias que permitan el monitoreo del buen funcionamiento del modelo.

Para futuros estudios se recomienda la inclusión de nuevas variables de negocio, fruto de la utilización y recopilación de datos adicionales, así como la exploración de otras alternativas de modelado.

Referencias

Agrawal et al. (2021). Selection of 51 predictors from 13,782 candidate multimodal features using machine learning improves coronary artery disease prediction. *Patterns*, 100364,.

Deepthi K., J. A. (2021). A deep learning ensemble approach to prioritize antiviral drugs against novel coronavirus SARS-CoV-2 for COVID-19 drug repurposing. *Deepthi K., Jereesh A.S. and Y. Liu*, 107945.

IArtificial.net. (2020, 09 10). *Las 7 Fases del Proceso de Machine Learning*. Retrieved from https://www.iartificial.net/fases-del-proceso-de-machine-learning/

IBM. (2021). *Conceptos básicos de ayuda de CRISP-DM*. Retrieved from https://www.ibm.com/docs/es/spss-modeler/SaaS?topic=dm-crisp-help-overview

Jing et al. (2020). A Machine Learning Approach to Management of Heart Failure Populations. *JACC: Heart Failure*, 578-587.

Lam, C. et al. (2021). Personalized stratification of hospitalization risk amidst COVID-19: A machine learning approach. *Health Policy and Technology*, 100554.

O.H. Salman, Z. T. (2021). A review on utilizing machine learning technology in the fields of electronic emergency triage and patient priority systems in telemedicine: Coherent taxonomy, motivations, open research challenges and recommendations for intelligent future work. *Computer Methods and Programs in Biomedicine*, 106357.

Reem, F. a. (2021). A novel Neutrosophic-based machine learning approach for maintenance prioritization in healthcare facilities. *Journal of Building Engineering, 42*, 102480.

Scikit-learn.org. (2021). *Linear and Quadratic Discriminant Analysis — scikit-learn 0.24.2 documentation.* Retrieved 07 2, 2021, from https://scikit-learn.org/stable/modules/lda_qda.html#lda-qda

Scikit-learn.org. (2021). *Naive Bayes — scikit-learn 0.24.2 documentation.* Retrieved 07 2, 2021, from https://scikit-learn.org/stable/modules/naive_bayes.html

Scikit-learn.org. (2021). *sklearn.linear_model.LogisticRegression — scikit-learn 0.24.2 documentation.* Retrieved 07 2, 2021, from https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html

Zhao et al. (2021). Social Determinants in Machine Learning Cardiovascular Disease Prediction Models: A Systematic Review. *American Journal of Preventive Medicine*, 596-605.