Informe de la fase N°.4 Modelos de Predicción para 2-Clases: Normal y Diabetes

Visión general:

Definir y caracterizar los mejores modelos predictivos para 2-clases, una clase “Normal” y otra clase “Diabetes” haciendo uso de un conjunto de datos de modelado el cual contiene datos clínicos, demográficos y de laboratorio ya acoplados y debidamente filtrados.

- Optimización y evaluación de modelos predictivos para 2-Clases: Normal, Diabetes
- Análisis de resultados
- Recomendaciones y próximos pasos

Miembros del equipo:

Incluir los nombres de los miembros del equipo que participaron tanto de la CCSS como del Adjudicado, Ingresar también los nombres de las partes relevantes, como interesados del negocio, personas involucradas en el trabajo y su rol, entre otros.

- Joaquín Zúñiga (Linchpin People)
- Eduardo Castro (Linchpin People)
- Kristina Ureña (Linchpin People)
- Eugenio Fernández (Linchpin People)
- Jason Diaz (Linchpin People)
- Roy Wong McClure (CCSS)
- Rosa María Matarrita Chaves (CCSS)
- Miriam León Solís (CCSS)
- Fernando Nassar Guier (CCSS)
- Zaira Cecilia Bastos Corella (CCSS)

Latam Training & Solutions S.A.
Cedula Jurídica 3-101-699202
Oficentro Ejecutivo La Sabana, Torre 6 Piso 7, oficina 26-801, Sabana Sur
Teléfono: (506) 4001-3282
Email: lataminfo@linchpinpeople.com

Aprendizajes

Se implementaron técnicas de importancia de variables, análisis no-supervisados y equilibrio de clases para poder caracterizar los modelos que mejor predicen y generalizan la predicción de las clases “Normal” y “Diabetes” usando variables simples y conocidas del proceso clínico y de laboratorio. No se desea utilizar la “Glucosa” como predictor ya que tiende a sesgar las capacidades predictivas del modelo y reduce su cobertura operativa al no ser una prueba que muchos tiene. Como veremos los algoritmos Gradient Boosting (GB), Random Forest (RF), Ada Boost(AB) y Suport Vector Machine (SVM-rbf) fueron los que mostraron mejore resultados.

Herramientas y métodos utilizados

Información sobre los productos, herramientas y servicios utilizados en la solución, así como métodos seguidos.

- Lenguaje de programación: Python 3.11, PyArrow, Pandas
- Azure Sypnase Analytics: PySpark, SkLearn
- Plataforma de análisis de calidad, exploración y preparación de datos desarrollada en las etapas anteriores: Jupyter Notebooks y métodos para ingeniería de datos
- Entorno de desarrollo: Azure Synapse Workspace
- Interfaz de desarrollo: Visual Studio
- Metodología de análisis y modelado CRISP-DM (https://www.sv-europe.com/crisp-dm-methodology/)

Latam Training & Solutions S.A.
Cedula Jurídica 3-101-699202
Oficentro Ejecutivo La Sabana, Torre 6 Piso 7, oficina 26-801, Sabana Sur
Teléfono: (506) 4001-3282
Email: lataminfo@linchpinpeople.com

Desafíos específicos

Se utilizaron técnicas de análisis de variables para obtener un subconjunto de componentes principales (predictores) mediante el análisis de componentes principales (PCA) para reducir la dimensionalidad del conjunto de datos de modelado que permita producir un modelo predictivo satisfactorio para las clases; Norma y Diabetes, maximizando el equilibrio entre precisión, sensibilidad y especificidad.

FASE 5. Modelos de Predicción para 2-Clases: Normal y Diabetes

Objetivos del negocio

La diabetes mellitus es enfermedad crónica que se caracteriza por un aumento en la cantidad de azúcar en la sangre y de la cual los tipos más frecuentes son: la diabetes gestacional que se presenta durante el embarazo, la diabetes mellitus tipo 1 que se presenta cuando el páncreas produce poca o ninguna insulina y la diabetes mellitus tipo 2 que es la más común y se da cuando el cuerpo se vuelve más resistente a la insulina (https://www.ministeriodesalud.go.cr/).

La Caja Costarricense de Seguro Social (CCSS) se ha propuesto como meta disminuir la incidencia de dicha enfermedad mediante la implementación de un sistema de prevención y detección temprana. Para lograr dicho objetivo es necesario consolidar datos clínicos y sociodemográficos históricos de pacientes con y sin diabetes. Dichos datos serán el insumo requerido por uno o varios modelos inteligentes basados en máquinas de aprendizaje y /o inteligencia artificial que tienen la capacidad de aprender del pasado para poder realizar una predicción futura por paciente. Dicha predicción tiene un margen de error que será definido y evaluado en su momento, pero que tiene el poder necesario de ayudar a que el sistema de salud pueda abordar un programa de prevención (modelo prescriptivo) por paciente.

Latam Training & Solutions S.A.
Cedula Jurídica 3-101-699202
Oficentro Ejecutivo La Sabana, Torre 6 Piso 7, oficina 26-801, Sabana Sur
Teléfono: (506) 4001-3282
Email: lataminfo@linchpinpeople.com

`

Evaluación de la situación

Se cuenta con los siguientes recursos on-premise y en la nube:

- Servidor de Machine Learning (ML) que cuenta con Python y Visual Studio Code instalado. La ingeniería de datos sigue realizándose en este servidor.
- Azure Synapse Analytics: PySpark y SkLearn,
- Herramientas o paquetes básicos y tradicionales para entendimiento, preparación y modelado de datos, tales como: PyArrow, Pandas, Scikit-Learn, Seaborn, Matplotlib y conectores a bases de datos.
- Los datos de los sistemas transaccionales de la CCSS deben ser extraídos y consolidados para que sean insumo para el proceso de modelaje y generación de modelo de ML.

Determinación de los objetivos del modelado

Específicamente, se busca obtener un modelo predictivo que permitan determinar la probabilidad de padecer o no Diabetes Mellitus Tipo 2. Este modelo permitirá a la CCSS reaccionar de manera proactiva ante enfermedades crónicas no transmisibles como la Diabetes Mellitus Tipo 2, entre otros, logrando un uso más eficiente de los recursos públicos.

Realización del plan de proyecto

Título: Desarrollo de un modelo de predicción para Diabetes Mellitus 2 en la Caja Costarricense de Seguro Social.

Descripción: Se utilizará datos históricos clínicos, sociodemográficos y de laboratorio de pacientes con y sin diabetes que servirán de insumo a máquinas de aprendizaje inteligentes (ML) y / o inteligencia artificial (AI) para que puedan realizar predicciones futuras basados en aprendizajes anteriores.

Latam Training & Solutions S.A.
Cedula Jurídica 3-101-699202
Oficentro Ejecutivo La Sabana, Torre 6 Piso 7, oficina 26-801, Sabana Sur
Teléfono: (506) 4001-3282
Email: lataminfo@linchpinpeople.com

1. Objetivo General: Desarrollar e implementar un modelo de predicción que permita determinar la probabilidad de padecer o no de diabetes mellitus tipo 2 para el sistema de salud pública de Costa Rica

2. **Objetivos Específicos**: A continuación, se incluyen los objetivos específicos abordados en este reporte técnico.

a. *Consolidar y habilitar el conjunto de datos clínicos y sociodemográficos históricos por paciente que contenga el mínimo de variables recomendadas para producir un modelo de predicción.*

b. *Definir los criterios de técnicos y de negocio mínimos requeridos para garantizar el éxito del modelo.*

c. *Entender los datos mediante la implementación de lógica de exploración, análisis, verificación y visualización de calidad de datos*

i. *Describir datos*

ii. *Explorar datos individual y colectivamente*

iii. *Verificar calidad de datos*

3. Explicación de los Datos:

Refiérase a la sección de entendimiento de los datos en Fase I.

4. Tipo(s) de Problema:

a. *Clustering (aprendizaje no supervisado)*

Latam Training & Solutions S.A.
Cedula Jurídica 3-101-699202
Oficentro Ejecutivo La Sabana, Torre 6 Piso 7, oficina 26-801, Sabana Sur
Teléfono: (506) 4001-3282
Email: lataminfo@linchpinpeople.com

b. *Predictivo (aprendizaje supervisado)*

c. *Regresión*

d. *Series de tiempo*

e. *Análisis simbólico*

f. *Otro (especifique): Análisis exploratorio y entendimiento de los datos*

5. Método(s) a Utilizar:

a. *Métodos exploratorios*

b. *Árboles de decisión*

c. *ACP*

d. *Máquinas vectoriales de soporte*

e. *Agrupación jerárquica*

f. *Bosques aleatorios*

g. *K-medias*

h. *Métodos de potenciación*

i. *K vecinos más cercanos*

j. *Regresión lineal*

k. *Regresión logística*

l. *Regresión RIDGE*

m. *Análisis discriminante lineal*

Latam Training & Solutions S.A.
Cedula Jurídica 3-101-699202
Oficentro Ejecutivo La Sabana, Torre 6 Piso 7, oficina 26-801, Sabana Sur
Teléfono: (506) 4001-3282
Email: lataminfo@linchpinpeople.com

n. *Regresión LASSO*

o. *Análisis discriminante cuadrático*

p. *Regresión Elastic Net*

q. *Métodos bayesianos*

r. *Series de tiempo*

s. *Redes neuronales*

t. *Métodos simbólicos*

u. *Otro (especifique): Visualización y análisis de datos*

6. Software(s) / Plataformas a utilizar:

a. Visual Studio Code
b. Azure Synapse Analytcis

b. Python

c. VPN Client

d. Microsoft Remote Desktop Connection

Latam Training & Solutions S.A.
Cedula Jurídica 3-101-699202
Oficentro Ejecutivo La Sabana, Torre 6 Piso 7, oficina 26-801, Sabana Sur
Teléfono: (506) 4001-3282
Email: lataminfo@linchpinpeople.com

`

Fuentes Primarias de Datos

Tipo de datos adquiridos: Se adquirió datos clínicos, demográficos y de laboratorio de pacientes con y sin diabetes mellitus que posee la Caja Costarricense de Seguro Social en sus sistemas.

Localización de fuente de datos:

- Servidor: 172.30.38.192
- Usuario: xxxxxxxx
- Clave: xxxxxxxxxx
- Base de datos: CCSS_DATAWAREHOUSE
- Tabla: vDSAselRegDiabetesMellitus

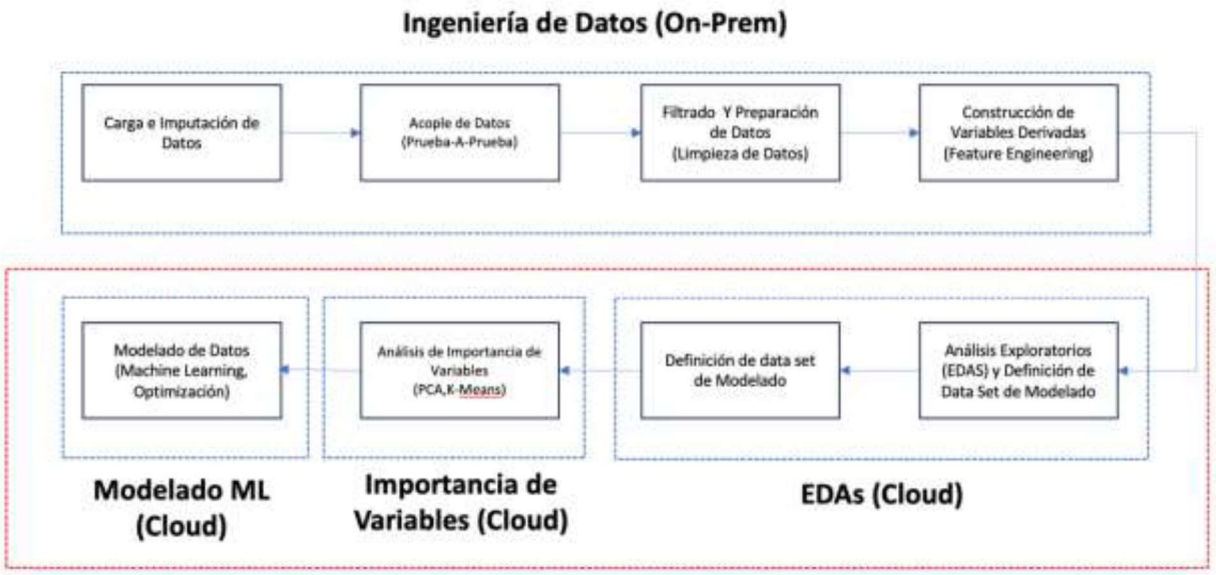
Recomendaciones y aprendizajes:

Durante esta fase de evaluación y optimización de modelos se logró determinar que el análisis de componentes principales como PCA nos permiten identificar un subconjunto de predictores con los cuales es posible construir un modelo predictor de 2-clases. Este proceso incluye técnicas de equilibrio de la variable objetivo con el objetivo de maximizar la sensibilidad y especificidad de los modelos.

Latam Training & Solutions S.A.
Cedula Jurídica 3-101-699202
Oficentro Ejecutivo La Sabana, Torre 6 Piso 7, oficina 26-801, Sabana Sur
Teléfono: (506) 4001-3282
Email: lataminfo@linchpinpeople.com

Flujo general de entendimiento, preparación, análisis y modelado de datos para diabetes tipo 2

A continuación, se muestra el flujo general de exploración, preparación, análisis y modelado de datos para diabetes. El cual nos proporciona los conjuntos de datos requeridos para la etapa de importancia de variable y modelado predictivo.



Latam Training & Solutions S.A.
Cedula Jurídica 3-101-699202
Oficentro Ejecutivo La Sabana, Torre 6 Piso 7, oficina 26-801, Sabana Sur
Teléfono: (506) 4001-3282
Email: lataminfo@linchpinpeople.com

`

Optimización y evaluación de modelos predictores para 2-Clases

En esta fase se avaluaron los resultados de 10 modelos predictores de predicción binaria, siguiendo las siguientes etapas:

E1. Definición del conjunto de datos de modelado: Aquí se aplican filtros descriptivos construidos en la etapa de ingeniería de atributos (Feature Engineering) para garantizar que los atributos predictores estén definidos cerca del momento de la "Fecha Consulta".
E2. Evaluación de algoritmos tomando en cuenta todos los predictores (**sin PCA**), y con dos metodologías:

a. Sin equilibrio
b. Con equilibrio

E3. Análisis de importancia de variables

E4. Evaluación de algoritmos tomando en cuenta solo las componentes principales definidas por el análisis de componentes importantes (**con PCA**), y con dos metodologías:

a. Sin equilibrio
b. Con equilibrio

Latam Training & Solutions S.A.
Cedula Jurídica 3-101-699202
Oficentro Ejecutivo La Sabana, Torre 6 Piso 7, oficina 26-801, Sabana Sur
Teléfono: (506) 4001-3282
Email: lataminfo@linchpinpeople.com

`

Definición del conjunto de datos de modelado

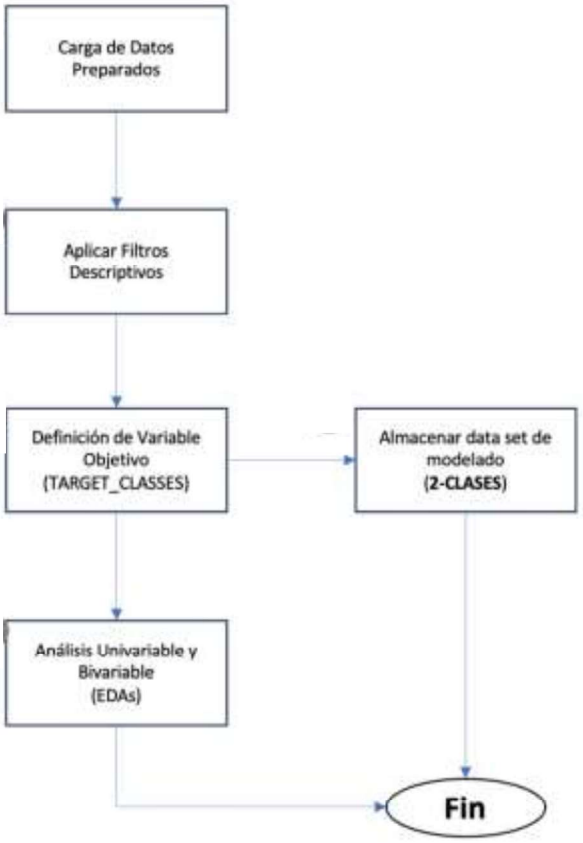
A continuación, se muestra el flujo de preparación de datos implementado usando Notebooks en Azure Synapse (izquierda), y los filtros principales implementado (derecha). Se consideran únicamente registros de pacientes que tengan resultados de laboratorio y comorbilidades con máximo 1 mes de lejanía, un historial clínico con 10 o más citas y que hayan estado al menos 1 año en la historia clínica. Luego se considera el subconjunto de pacientes que solo tienen registros sin diabetes tipo 2 (“Normales”), y al subconjunto de pacientes que solo tienen registros con diabetes (“Diabetes”). Los pacientes que contienen registros mixtos (unos normales y otros con diabetes) son excluidos para reducir la variabilidad de los datos predictores (revisar Notebooks EDAs en Azure Synapse).

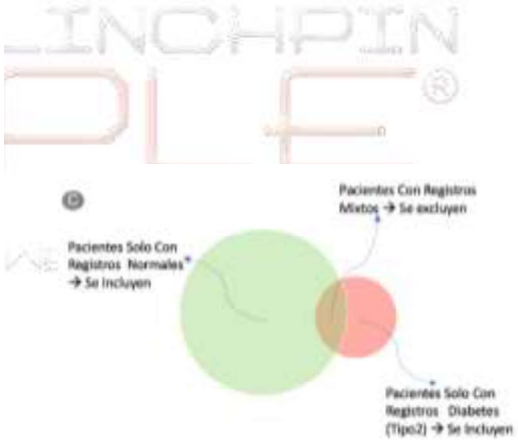
Notas:

- Los datos preparados (A) fueron construidos y definidos en la etapa de ingeniería de datos desarrollada y documentada en los informes anteriores.
- Los resultados de los EDAs no se agregan en este informe técnico ya que han sido revisados y documentados en los informes anteriores. Dado que es un proceso reiterativo evaluación y de mejora continua, favor referirse a los notebooks de EDAs directamente en la plataforma de AZ Synapse.
- Para el conjunto de datos de modelado de 2-Clases la variable objetivo “TARGET_CLASES” se define usando la variable derivada “dm_tipo2” (revisar Notebooks EDAs en Azure Synapse).

Latam Training & Solutions S.A.
Cedula Jurídica 3-101-699202
Oficentro Ejecutivo La Sabana, Torre 6 Piso 7, oficina 26-801, Sabana Sur
Teléfono: (506) 4001-3282
Email: lataminfo@linchpinpeople.com

El flujo seguido para la construcción del conjunto de datos para modelado se muestra en la siguiente imagen.





Latam Training & Solutions S.A.
Cedula Jurídica 3-101-699202
Oficentro Ejecutivo La Sabana, Torre 6 Piso 7, oficina 26-801, Sabana Sur
Teléfono: (506) 4001-3282
Email: lataminfo@linchpinpeople.com

`

Evaluación de algoritmos sin PCA

En esta etapa se realiza una evaluación de 10 distintos modelos, utilizando 5 metodologías distintas y considerando todas las columnas del conjunto de datos. Los modelos evaluados son construidos utilizando son los siguientes algoritmos:

- GB. Gradient Boost
- RF. Random Forest
- LR. Logistic Regresion
- KNN. Nearest Neighbors
- DT. Decision Trees
- NB. GaussianNB
- AB. Ada Boost
- QDA. Quadratic Discriminant Analysis
- MLP. Multi-Layer Perceptron
- SVM-Linear. Support Vector Machine Linear.
- SVM-RBF. Support Vector Machine RBF.

Cada uno de estos modelos fue evaluado según las siguientes estrategias:

- Estrategia 1. Evaluación de modelos sin equilibrio, todos los atributos, **todos los registros** por paciente (con y sin diabetes)
- Estrategia 2. Evaluación de modelos sin equilibrio, todos los atributos, **registro más reciente** por paciente (con y sin diabetes)
- Estrategia 3. Evaluación de modelos con equilibrio usando algoritmo de submuestreo aleatorio **RUS**[1]

[1] https://imbalanced-learn.org/stable/references/generated/imblearn.under_sampling.RandomUnderSampler.html

Latam Training & Solutions S.A.
Cedula Jurídica 3-101-699202
Oficentro Ejecutivo La Sabana, Torre 6 Piso 7, oficina 26-801, Sabana Sur
Teléfono: (506) 4001-3282
Email: lataminfo@linchpinpeople.com

- Estrategia 4. Evaluación de modelos con equilibrio usando algoritmo de submuestreo **NearMiss**[2]
- Estrategia 5. Evaluación de modelos con equilibrio usando algoritmo de submuestreo **TomekLinks**[3]

Tal y como se muestra en la siguiente imagen, las primeras dos estrategias (1 y 2) muestra buenos resultados en términos de precisión, pero mal balance entre sensibilidad y especificidad. Mientras todos los modelos son buenos prediciendo pacientes que pertenecen a la clase "**Normal**", ninguno es bueno prediciendo la clases "**Diabética**", la cual es el objetivo principal del proyecto. Esto se debe fundamentalmente a que las clases a predecir están desbalanceadas (la cantidad de registros normales es mucho mayor que la cantidad de registros con diabetes).

METHODOLOGY	PERFORMANCE PARAM	GB	RF	LR	KNN	DT	NB	AB	QDA	MLP	SVM-Linear
Estrategia 1. wo_PCA_wo_Hypertuning_All_Filtered_ Records (564K)	Accuracy	93%	93%	93%	93%	88%	12%	93%	16%	93%	93%
	Macro-Avg Precision	71%	73%	58%	61%	56%	53%	65%	51%	47%	47%
	Macro-Avg Recall	50%	50%	50%	52%	57%	53%	50%	52%	50%	50%
	Macro-Avg F1-score	49%	49%	49%	53%	56%	12%	49%	16%	48%	48%
	Weighted-Avg Precision	90%	91%	88%	89%	89%	92%	89%	89%	87%	87%
	Weighted-Avg Recall	93%	93%	93%	93%	88%	12%	93%	16%	93%	93%
	Weighted-Avg F1-score	90%	90%	90%	90%	89%	12%	90%	19%	90%	90%
	Sensitivity (TPR) --> NORMAL	100%	100%	100%	99%	93%	6%	100%	11%	100%	100%
	Sensitivity (TPR) --> DIABETES	0%	1%	0%	6%	20%	98%	1%	92%	0%	0%
	Specificity(TNR) --> NORMAL	0%	1%	0%	6%	20%	98%	1%	92%	0%	0%
	Specificity(TNR) --> DIABETES	100%	100%	100%	99%	93%	6%	100%	11%	100%	100%
Estrategia 2. wo_PCA_wo_Hypertuning_Last_Filtered _Record (325K)	Accuracy	95%	95%	95%	94%	91%	11%	95%	11%	95%	95%
	Macro-Avg Precision	77%	61%	58%	59%	56%	52%	75%	51%	47%	47%
	Macro-Avg Recall	50%	50%	50%	51%	56%	53%	50%	55%	50%	50%
	Macro-Avg F1-score	49%	49%	49%	51%	56%	11%	49%	35%	49%	49%
	Weighted-Avg Precision	93%	91%	91%	91%	91%	94%	93%	91%	90%	90%
	Weighted-Avg Recall	95%	95%	95%	94%	91%	11%	95%	44%	95%	95%
	Weighted-Avg F1-score	92%	92%	92%	92%	91%	12%	92%	56%	92%	92%
	Sensitivity (TPR) --> NORMAL	100%	100%	100%	99%	95%	7%	100%	42%	100%	100%
	Sensitivity (TPR) --> DIABETES	0%	0%	0%	3%	17%	99%	1%	67%	0%	0%
	Specificity(TNR) --> NORMAL	0%	0%	0%	3%	17%	99%	1%	67%	0%	0%
	Specificity(TNR) --> DIABETES	100%	100%	100%	100%	95%	7%	100%	42%	100%	100%

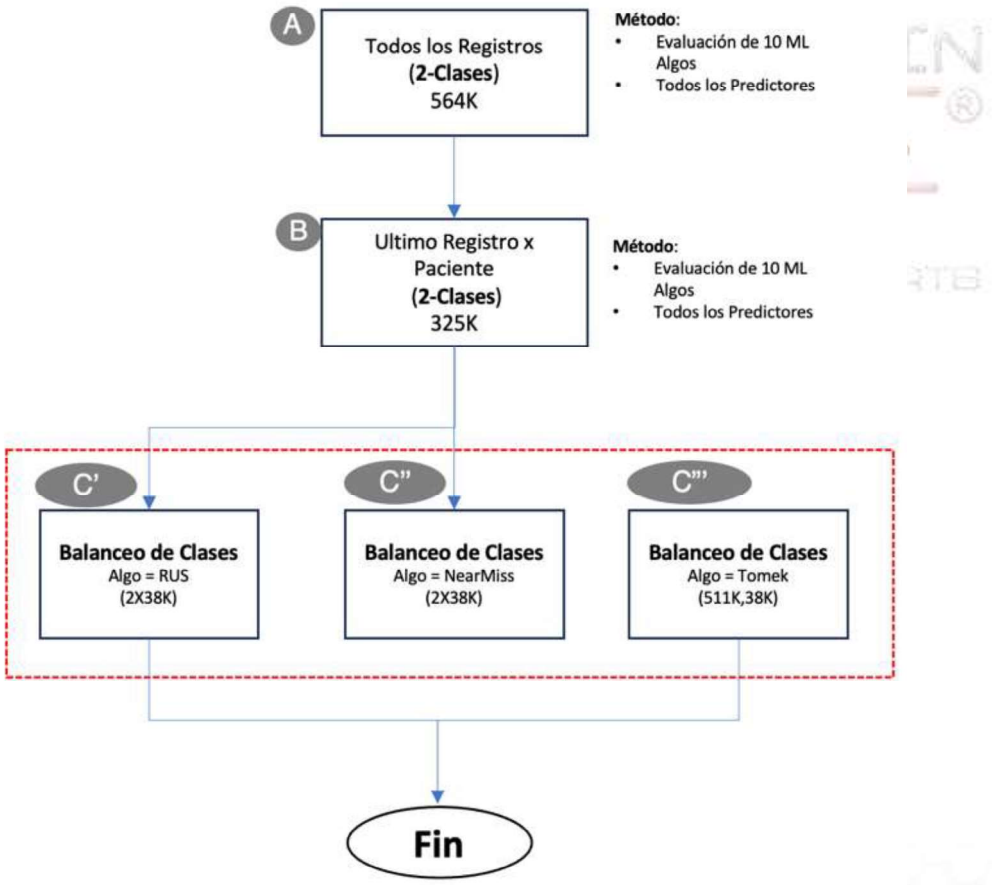
Para obtener mejores resultados en el aprendizaje de maquina (ML Modeling) se recomienda implementar algoritmos de equilibrio. Debido a que se cuenta con un

[2] https://imbalanced-learn.org/stable/references/generated/imbleam.under_sampling.NearMiss.html)
[3] https://imbalanced-learn.org/stable/references/generated/imbleam.under_sampling.TomekLinks.html

Latam Training & Solutions S.A.
Cedula Jurídica 3-101-699202
Oficentro Ejecutivo La Sabana, Torre 6 Piso 7, oficina 26-801, Sabana Sur
Teléfono: (506) 4001-3282
Email: lataminfo@linchpinpeople.com

volumen importante de registros con la clase “Diabetes”, se recomienda implementar metodologías de “submuestreo”. Existen varios algoritmos en este campo y se implementaron los tres más utilizados en la industria los cuales son RUS, NearMiss y TomekLinks.

La siguiente figura muestra el flujo aplicado en la etapa de equilibrio de clases. Los algoritmos de equilibrio se aplican a los conjuntos de datos utilizados como entrada para el entrenamiento y evaluación de modelos.



Una vez aplicados los métodos de equilibrio de datos, se determinó que los algoritmos de equilibrio RUS y NearMiss logran balancear y maximizar la sensibilidad y la especificidad de los modelos.

Latam Training & Solutions S.A.
Cedula Jurídica 3-101-699202
Oficentro Ejecutivo La Sabana, Torre 6 Piso 7, oficina 26-801, Sabana Sur
Teléfono: (506) 4001-3282
Email: lataminfo@linchpinpeople.com

Específicamente, RUS brinda mejor capacidad de predecir pacientes que pertenecen a la clase “Diabetes” con rangos entre 76% y 86% de precisión, mientras que NearMiss brinda mejor capacidad de predecir pacientes que pertenecen a la clase “Normal” con rangos entre 81% y 88% de precisión.

La siguiente tabla es un resumen de los resultados de evaluación de precisión, sensibilidad y la especificidad de los modelos utilizando las técnicas de balance de clases, en la cual se puede apreciar el incremento que tienen los algoritmos en sensibilidad una vez aplicadas las técnicas de equilibrio de datos.

METHODOLOGY	PERFORMANCE PARAM	GB	RF	LR	KNN	DT	NB	AB	QDA	MLP	SVM-Linear	SVM-RBF
Estrategia 3. wo_PCA__UnderSampling_RUS (2x38k)	Accuracy	**76%**	**77%**	71%	71%	68%	52%	**75%**	50%	72%	**72%**	**75%**
	Macro-Avg Precision	**77%**	**78%**	71%	71%	68%	66%	**76%**	50%	73%	**72%**	**76%**
	Macro-Avg Recall	**76%**	**77%**	71%	71%	68%	52%	**75%**	50%	72%	**72%**	**75%**
	Macro-Avg F1-score	**76%**	**77%**	71%	71%	68%	39%	**75%**	49%	72%	**72%**	**75%**
	Weighted-Avg Precision	**77%**	**78%**	71%	71%	68%	66%	**76%**	50%	73%	**72%**	**76%**
	Weighted-Avg Recall	**76%**	**77%**	71%	71%	68%	52%	**75%**	50%	72%	**72%**	**75%**
	Weighted-Avg F1-score	**76%**	**77%**	71%	71%	68%	39%	**75%**	49%	72%	**72%**	**75%**
	Sensitivity (TPR) --> NORMAL	**67%**	**69%**	70%	66%	69%	6%	**66%**	41%	66%	**68%**	**65%**
	Sensitivity (TPR) --> DIABETES	**86%**	**85%**	72%	75%	67%	99%	**84%**	58%	78%	**76%**	**85%**
	Specificity(TNR) --> NORMAL	**86%**	**85%**	72%	75%	67%	99%	**84%**	58%	78%	**76%**	**85%**
	Specificity(TNR) --> DIABETES	**67%**	**69%**	70%	66%	69%	6%	**66%**	41%	66%	**68%**	**65%**
Estrategia 4. wo_PCA_UnderSampling_NearMiss(2x38K)	Accuracy	**80%**	**79%**	73%	72%	72%	54%	**78%**	50%	73%	**74%**	**80%**
	Macro-Avg Precision	**80%**	**79%**	73%	72%	72%	76%	**78%**	69%	74%	**74%**	**81%**
	Macro-Avg Recall	**80%**	**79%**	73%	72%	72%	53%	**78%**	50%	73%	**73%**	**80%**
	Macro-Avg F1-score	**80%**	**79%**	73%	72%	72%	40%	**78%**	33%	73%	**73%**	**80%**
	Weighted-Avg Precision	**80%**	**79%**	73%	72%	72%	76%	**78%**	69%	74%	**74%**	**81%**
	Weighted-Avg Recall	**80%**	**79%**	73%	72%	72%	64%	**78%**	50%	73%	**73%**	**80%**
	Weighted-Avg F1-score	**80%**	**79%**	73%	72%	72%	41%	**78%**	33%	73%	**73%**	**80%**
	Sensitivity (TPR) --> NORMAL	**85%**	**83%**	78%	84%	72%	100%	**82%**	0%	80%	**81%**	**88%**
	Sensitivity (TPR) --> DIABETES	**74%**	**75%**	69%	61%	71%	6%	**74%**	100%	67%	**66%**	**72%**
	Specificity(TNR) --> NORMAL	**74%**	**75%**	69%	61%	71%	6%	**74%**	100%	67%	**66%**	**72%**
	Specificity(TNR) --> DIABETES	**86%**	**83%**	78%	84%	72%	100%	**82%**	0%	83%	**81%**	**88%**
Estrategia 5. wo_PCA_UnderSampling_Tomek({'NORMAL': 511K, 'DIABETES': 38K})	Accuracy	93%	93%	93%	93%	88%	13%	93%	19%	93%	93%	N/A
	Macro-Avg Precision	76%	78%	59%	65%	57%	53%	71%	52%	47%	47%	N/A
	Macro-Avg Recall	50%	51%	50%	54%	58%	52%	51%	53%	50%	50%	N/A
	Macro-Avg F1-score	49%	50%	49%	55%	57%	13%	49%	19%	48%	48%	N/A
	Weighted-Avg Precision	91%	91%	88%	89%	89%	92%	90%	90%	87%	87%	N/A
	Weighted-Avg Recall	93%	93%	93%	93%	88%	13%	93%	19%	93%	93%	N/A
	Weighted-Avg F1-score	90%	90%	90%	90%	89%	12%	90%	23%	90%	90%	N/A
	Sensitivity (TPR) --> NORMAL	100%	100%	100%	99%	93%	6%	100%	13%	100%	100%	N/A
	Sensitivity (TPR) --> DIABETES	1%	2%	1%	9%	22%	98%	1%	93%	0%	0%	N/A
	Specificity(TNR) --> NORMAL	1%	2%	1%	9%	22%	98%	1%	93%	0%	0%	N/A
	Specificity(TNR) --> DIABETES	100%	100%	100%	99%	93%	6%	100%	13%	100%	100%	N/A

Con base en los resultados de los experimentos realizados con las distintas estrategias de equilibrio de datos, se ordenan los resultados de evaluación de modelos según la

Latam Training & Solutions S.A.
Cedula Jurídica 3-101-699202
Oficentro Ejecutivo La Sabana, Torre 6 Piso 7, oficina 26-801, Sabana Sur
Teléfono: (506) 4001-3282
Email: lataminfo@linchpinpeople.com

precisión de estos y la sensibilidad con la predicción de la clase DIABETES, en las siguientes tablas se resumen dichos resultados.

ESTRATEGIA	ALGORITMO	PRECISIÓN	SENSIBILIDAD (DIABETES)
Estrategia 3. wo_PCA__UnderSampling_RUS	RF	77%	85%
	GB	76%	86%
	SVM-RBF	75%	85%
	AB	75%	84%
	MLP	72%	78%
	SVM-Linear	72%	76%
	KNN	71%	75%
	LR	71%	72%
	DT	68%	67%
	NB	52%	99%
	QDA	50%	58%

ESTRATEGIA	ALGORITMO	PRECISIÓN	SENSIBILIDAD (DIABETES)
Estrategia 4. wo_PCA_UnderSampling_NearMiss	GB	80%	74%
	SVM-RBF	80%	72%
	RF	79%	75%
	AB	78%	74%
	SVM-Linear	74%	66%
	LR	73%	69%
	MLP	73%	67%
	DT	72%	71%
	KNN	72%	61%
	NB	54%	6%
	QDA	50%	100%

Latam Training & Solutions S.A.
Cedula Jurídica 3-101-699202
Oficentro Ejecutivo La Sabana, Torre 6 Piso 7, oficina 26-801, Sabana Sur
Teléfono: (506) 4001-3282
Email: lataminfo@linchpinpeople.com

`

ESTRATEGIA	ALGORITMO	PRECISIÓN	SENSIBILIDAD (DIABETES)
Estrategia 5. wo_PCA_UnderSampling_Tomek	KNN	93%	9%
	RF	93%	2%
	GB	93%	1%
	LR	93%	1%
	AB	93%	1%
	MLP	93%	0%
	SVM-Linear	93%	0%
	DT	88%	22%
	QDA	19%	93%
	NB	13%	98%
	SVM-RBF	0%	0%

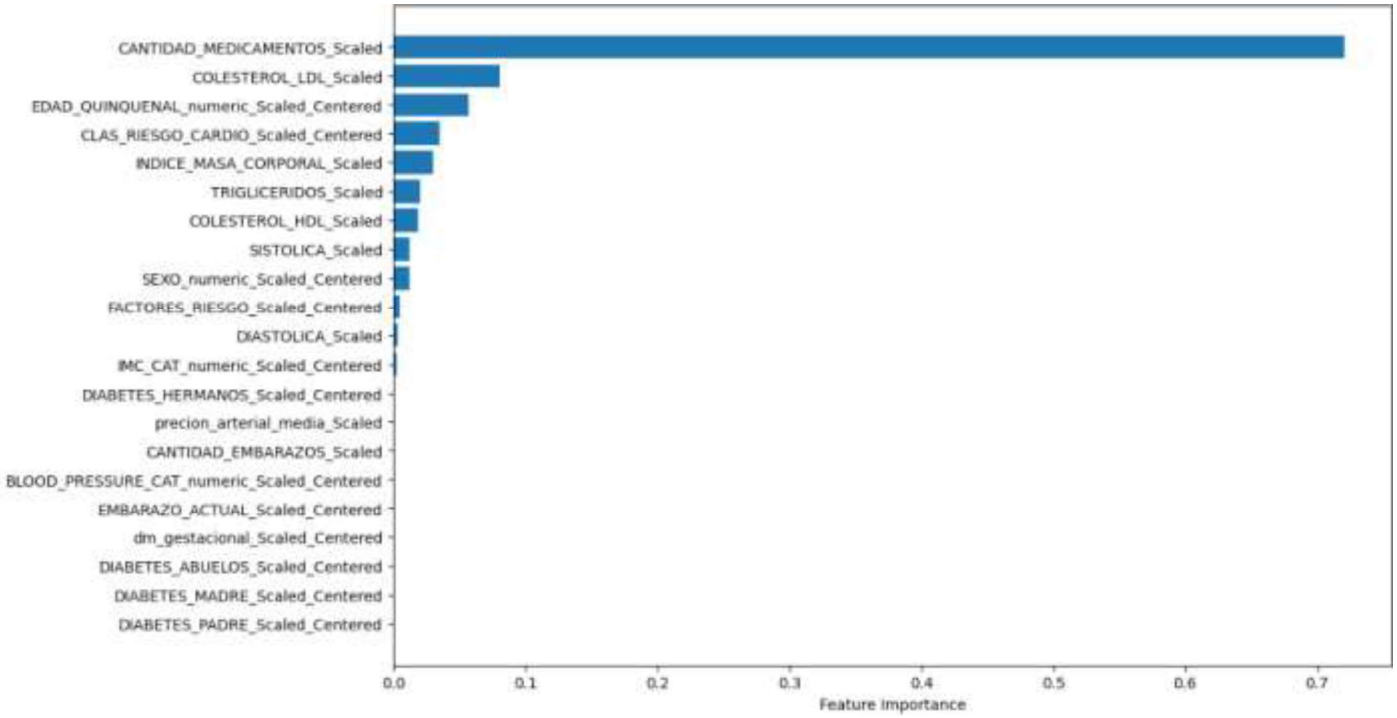
Con base en los resultados anteriores se determina que los algoritmos que tienen mejor precisión y sensibilidad son Gradient Boosting (GB), Random Forest (RF) y Support Vector Machine (SVM).

Una vez determinados los modelos con mejor precisión se procede a calcular los coeficientes de cada uno de los modelos basados en dichos algoritmos.

Latam Training & Solutions S.A.
Cedula Jurídica 3-101-699202
Oficentro Ejecutivo La Sabana, Torre 6 Piso 7, oficina 26-801, Sabana Sur
Teléfono: (506) 4001-3282
Email: lataminfo@linchpinpeople.com

Coeficientes para algoritmo Gradient Boost GB

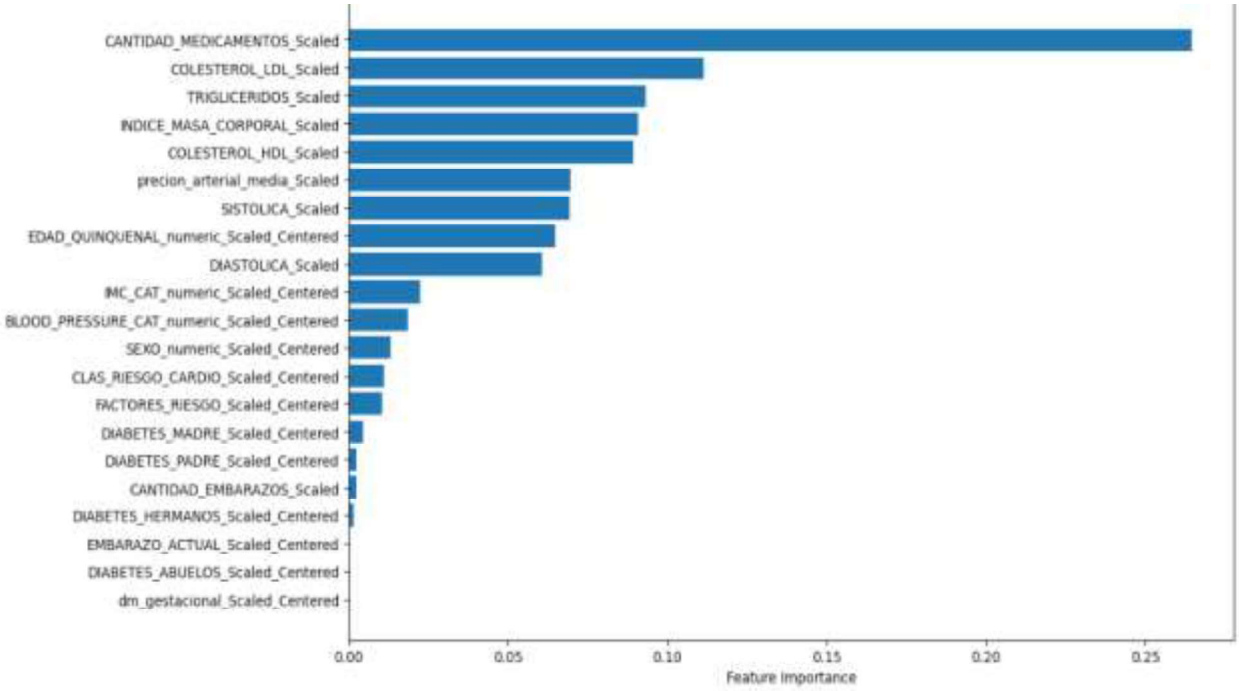
El siguiente gráfico muestra la lista de predictores ordenados por importancia para el algoritmo GB:



Latam Training & Solutions S.A.
Cedula Jurídica 3-101-699202
Oficentro Ejecutivo La Sabana, Torre 6 Piso 7, oficina 26-801, Sabana Sur
Teléfono: (506) 4001-3282
Email: lataminfo@linchpinpeople.com

Coeficientes para algoritmo Random Forest RF

El siguiente gráfico muestra la lista de predictores ordenados por importancia para el algoritmo RF:



De los gráficos anteriores se puede concluir que “Cantidad de Medicamentos”, “Colesterol LDL”, “Triglicéridos”, “Índice de Masa Corporal” y “Edad Quinquenal” tienden a tener más importancia en general. Sin embargo, Random Forest muestra una distribución más balanceada que los otros algoritmos.

Latam Training & Solutions S.A.
Cedula Jurídica 3-101-699202
Oficentro Ejecutivo La Sabana, Torre 6 Piso 7, oficina 26-801, Sabana Sur
Teléfono: (506) 4001-3282
Email: lataminfo@linchpinpeople.com

Evaluación de modelos considerando análisis de importancia de variables

La evaluación de modelos anterior considera 21 predictores, lo cual incrementa la complejidad y operacionalización del modelo de predicción.

Por lo tanto, en esta sección se incluyen los resultados de los experimentos de los modelos utilizando menos predictores los cuales son el resultado de la aplicación de la técnica de PCA (Análisis de Componentes Principales).
El objetivo del análisis de componentes principales (reducción de dimensionalidad) es encontrar un subconjunto de predictores que puedan explicar la variabilidad de los datos con un grado de confianza alto o aceptable (>90%). Este subconjunto de predictores (o componentes principales) debería ser suficiente para producir un modelo predictor.

Dado que el conjunto de datos de modelado cuenta con predictores numéricos y predictores categóricos, es requerido implementar la metodología de análisis de factores para datos mixtos, conocida en inglés como FAMD[4], para poder aplicar adecuadamente el análisis de componentes principales PCA. Los siguientes son los pasos principales abordados en la metodología FAMD:

1. Estandarizar variables numéricas (ejemplo: z-score)
2. Estandarizar variables categóricas:
    a. Aplicar decodificación “One-Hot” para aquellas variables categóricas y que requieran ser procesadas como numéricas en el análisis PCA.
    b. Estandarizar variables “One-Hoted”
3. Aplicar algoritmo PCA con distintos “solvers”, con el objetivo de encontrar el

[4] https://towardsdatascience.com/famd-how-to-generalize-pca-to-categorical-and-numerical-data-2ddbeb2b9210

Latam Training & Solutions S.A.
Cedula Jurídica 3-101-699202
Oficentro Ejecutivo La Sabana, Torre 6 Piso 7, oficina 26-801, Sabana Sur
Teléfono: (506) 4001-3282
Email: lataminfo@linchpinpeople.com

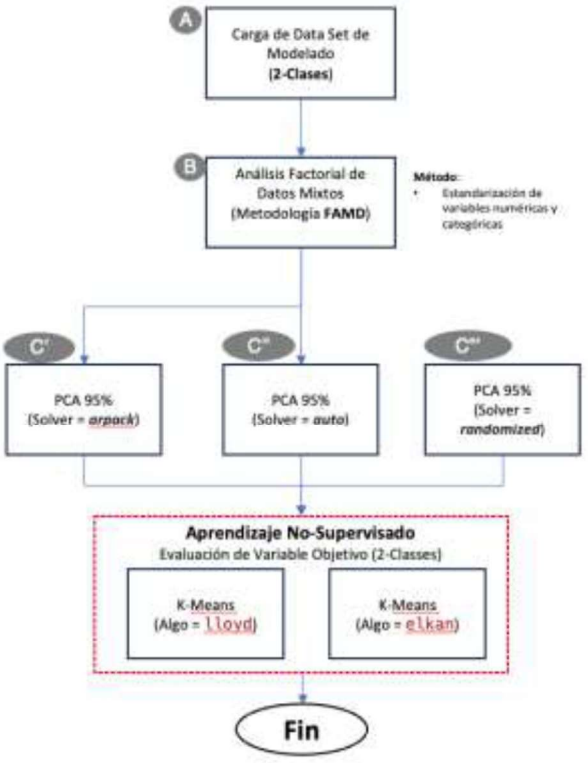
subconjunto de componentes principales común aplicando aleatoriedad y distintos métodos de resolución.

4. Evaluación no-supervisada de variable objetivo, con el objetivo de encontrar aquellos registros que presentan una baja consistencia a la clase que podrían pertenecer.
5. Analizar resultados.

Nota:

- Los detalles sobre los análisis PCA fueron documentados en el informe anterior.

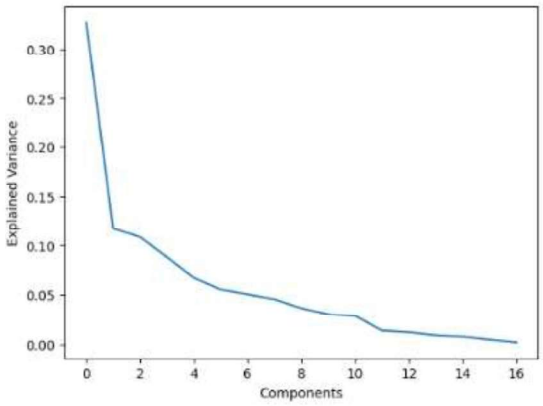
El siguiente diagrama muestra el flujo seguido para implementar el PCA.



Latam Training & Solutions S.A.
Cedula Jurídica 3-101-699202
Oficentro Ejecutivo La Sabana, Torre 6 Piso 7, oficina 26-801, Sabana Sur
Teléfono: (506) 4001-3282
Email: lataminfo@linchpinpeople.com

Una vez aplicado el PCA da como resultado las siguientes variables.

B



C

```
['SEXO_numeric_Scaled_Centered',
'FACTORES_RIESGO_Scaled_Centered',
'EDAD_QUINQUENAL_numeric_Scaled_Centered',
'DIABETES_MADRE_Scaled_Centered',
'IMC_CAT_numeric_Scaled_Centered',
'TRIGLICERIDOS_Scaled',
'BLOOD_PRESSURE_CAT_numeric_Scaled_Centered',
'CANTIDAD_MEDICAMENTOS_Scaled',
'DIABETES_PADRE_Scaled_Centered',
'COLESTEROL_LDL_Scaled']
```

El análisis FAMD (generalización de PCA) demuestra que es posible caracterizar aproximadamente el 95% de la variabilidad de los datos con un subconjunto de 10 a 12 componentes principales (y no 21 predictores, como en la etapa anterior). Los componentes definidos en “C” serán los predictores a utilizar.

Latam Training & Solutions S.A.
Cedula Jurídica 3-101-699202
Oficentro Ejecutivo La Sabana, Torre 6 Piso 7, oficina 26-801, Sabana Sur
Teléfono: (506) 4001-3282
Email: lataminfo@linchpinpeople.com

Evaluación de algoritmos con PCA

En esta etapa se realiza una evaluación de 10 distintos modelos utilizando distintas estrategias y considerando únicamente las variables identificadas con la aplicación del PCA. Los modelos evaluados son construidos utilizando son los siguientes algoritmos:

- GB. Gradient Boost
- RF. Random Forest
- LR. Logistic Regresion
- KNN. Nearest Neighbors
- DT. Decision Trees
- NB. GaussianNB
- AB. Ada Boost
- QDA. Quadratic Discriminant Analysis
- MLP. Multi-Layer Perceptron
- SVM-Linear. Support Vector Machine Linear.
- SVM-RBF. Support Vector Machine RBF.

Cada uno de estos modelos fue evaluado según las siguientes estrategias:

- Estrategia 1. Evaluación de modelos sin equilibrio, todos los atributos, **todos los registros** por paciente (con y sin diabetes)
- Estrategia 2. Evaluación de modelos sin equilibrio, todos los atributos, **registro más reciente** por paciente (con y sin diabetes)
- Estrategia 3. Evaluación de modelos con equilibrio usando algoritmo de submuestreo aleatorio **RUS**[5]

[5] https://imbalanced-learn.org/stable/references/generated/imblearn.under_sampling.RandomUnderSampler.html

Latam Training & Solutions S.A.
Cedula Jurídica 3-101-699202
Oficentro Ejecutivo La Sabana, Torre 6 Piso 7, oficina 26-801, Sabana Sur
Teléfono: (506) 4001-3282
Email: lataminfo@linchpinpeople.com

- Estrategia 4. Evaluación de modelos con equilibrio usando algoritmo de submuestreo **NearMiss**[6]

Tal y como se muestra en la siguiente imagen, las primeras dos estrategias sin equilibrio de datos (1 y 2) mostraron nuevamente buenos resultados en términos de precisión, pero mal balance entre sensibilidad y especificidad. Estos modelos son buenos prediciendo pacientes que pertenecen a la clase "**Normal**", pero no la clase "**Diabética**".

METHODOLOGY	PERFORMANCE PARAM	GB	RF	LR	KNN	DT	NB	AB	QDA	MLP	SVM-Linear	SVM-RBF
Estrategia 1. wo_PCA_wo_Hypertuning_All_Filtered_Records	Accuracy	93%	93%	93%	93%	88%	12%	93%	16%	93%	93%	0%
	Macro-Avg Precision	71%	73%	58%	61%	56%	53%	65%	51%	47%	47%	0%
	Macro-Avg Recall	50%	50%	50%	52%	57%	53%	50%	52%	50%	50%	0%
	Macro-Avg F1-score	49%	49%	49%	53%	56%	12%	49%	16%	48%	48%	0%
	Weighted-Avg Precision	90%	91%	88%	89%	89%	92%	89%	89%	87%	87%	0%
	Weighted-Avg Recall	93%	93%	93%	93%	88%	12%	93%	16%	93%	93%	0%
	Weighted-Avg F1-score	90%	90%	90%	90%	89%	12%	90%	19%	90%	90%	0%
	Sensitivity (TPR) --> NORMAL	100%	100%	100%	99%	93%	6%	100%	11%	100%	100%	0%
	Sensitivity (TPR) --> DIABETES	0%	1%	0%	6%	20%	98%	1%	92%	0%	0%	0%
	Specificity(TNR) --> NORMAL	0%	1%	0%	6%	20%	98%	1%	92%	0%	0%	0%
	Specificity(TNR) --> DIABETES	100%	100%	100%	99%	93%	6%	100%	11%	100%	100%	0%
Estrategia 2. wo_PCA_wo_Hypertuning_Last_Filtered_Record	Accuracy	95%	95%	95%	94%	91%	11%	95%	11%	95%	95%	0%
	Macro-Avg Precision	77%	61%	58%	59%	56%	52%	75%	51%	47%	47%	0%
	Macro-Avg Recall	50%	50%	50%	51%	56%	53%	50%	55%	50%	50%	0%
	Macro-Avg F1-score	49%	49%	49%	51%	56%	11%	49%	35%	49%	49%	0%
	Weighted-Avg Precision	93%	91%	91%	91%	91%	94%	93%	91%	90%	90%	0%
	Weighted-Avg Recall	95%	95%	95%	94%	91%	11%	95%	44%	95%	95%	0%
	Weighted-Avg F1-score	92%	92%	92%	92%	91%	12%	92%	56%	92%	92%	0%
	Sensitivity (TPR) --> NORMAL	100%	100%	100%	99%	95%	7%	100%	42%	100%	100%	0%
	Sensitivity (TPR) --> DIABETES	0%	0%	0%	3%	17%	99%	1%	67%	0%	0%	0%
	Specificity(TNR) --> NORMAL	0%	0%	0%	3%	17%	99%	1%	67%	0%	0%	0%
	Specificity(TNR) --> DIABETES	100%	100%	100%	100%	95%	7%	100%	42%	100%	100%	0%

Posteriormente se procede a realizar los experimentos con los mismos algoritmos, pero aplicando distintas técnicas de equilibrio de datos. Una vez finalizados los experimentos se determinó que los algoritmos de equilibrio (submuestreo) RUS y NearMiss nuevamente logran balancear y maximizar la sensibilidad y la especificidad de los modelos en general.

[6] https://imbalanced-learn.org/stable/references/generated/imblearn.under_sampling.NearMiss.html)

Latam Training & Solutions S.A.
Cedula Jurídica 3-101-699202
Oficentro Ejecutivo La Sabana, Torre 6 Piso 7, oficina 26-801, Sabana Sur
Teléfono: (506) 4001-3282
Email: lataminfo@linchpinpeople.com

RUS nos brinda mejor capacidad de predecir pacientes que pertenecen a la clase "Diabetes" con 75% de precisión con RF y sensibilidad de 81%.

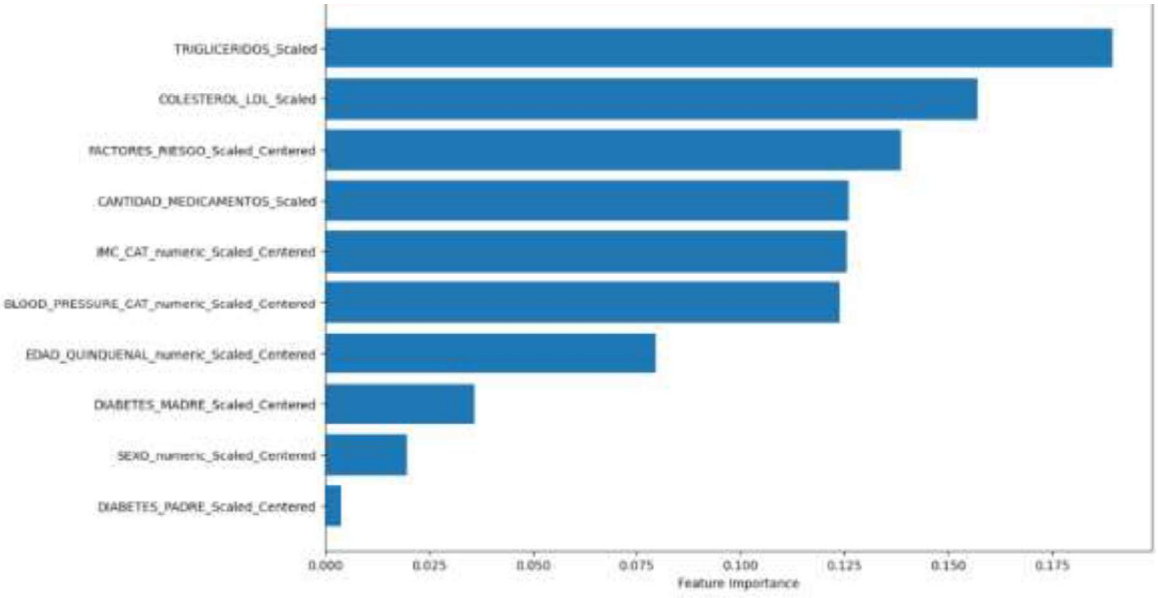
METHODOLOGY	PERFORMANCE PARAM	GB	RF	LR	KNN	DT	NB	AB	QDA	MLP	SVM-Linear	SVM-RBF
Estrategia 3. wo_PCA__UnderSampling_RUS	Accuracy	74%	75%	71%	71%	68%	52%	74%	50%	72%	71%	73%
	Macro-Avg Precision	75%	75%	71%	71%	68%	66%	75%	50%	73%	71%	75%
	Macro-Avg Recall	74%	75%	71%	71%	68%	52%	74%	50%	72%	71%	74%
	Macro-Avg F1-score	74%	74%	71%	71%	68%	39%	74%	49%	72%	71%	73%
	Weighted-Avg Precision	75%	75%	71%	71%	68%	66%	75%	50%	73%	71%	75%
	Weighted-Avg Recall	74%	75%	71%	71%	68%	52%	74%	50%	72%	71%	73%
	Weighted-Avg F1-score	74%	74%	71%	71%	68%	39%	74%	49%	72%	71%	73%
	Sensitivity (TPR) --> NORMAL	66%	69%	70%	66%	69%	6%	65%	41%	66%	67%	63%
	Sensitivity (TPR) --> DIABETES	83%	81%	72%	75%	67%	99%	83%	58%	78%	74%	84%
	Specificity(TNR) --> NORMAL	83%	81%	72%	75%	67%	99%	83%	58%	78%	74%	84%
	Specificity(TNR) --> DIABETES	66%	69%	70%	66%	69%	6%	65%	41%	66%	67%	63%
Estrategia 4. wo_PCA__UnderSampling_Near Miss	Accuracy	79%	77%	73%	72%	72%	54%	78%	50%	73%	74%	79%
	Macro-Avg Precision	80%	77%	73%	72%	72%	76%	79%	69%	74%	74%	80%
	Macro-Avg Recall	79%	77%	73%	72%	72%	53%	78%	50%	73%	74%	79%
	Macro-Avg F1-score	79%	77%	73%	72%	72%	40%	78%	33%	73%	74%	79%
	Weighted-Avg Precision	80%	77%	73%	72%	72%	76%	79%	69%	74%	74%	80%
	Weighted-Avg Recall	79%	77%	73%	72%	72%	64%	78%	50%	73%	74%	79%
	Weighted-Avg F1-score	79%	77%	73%	72%	72%	41%	78%	33%	73%	74%	79%
	Sensitivity (TPR) --> NORMAL	85%	81%	78%	84%	72%	100%	84%	0%	80%	81%	89%
	Sensitivity (TPR) --> DIABETES	73%	72%	69%	61%	71%	6%	72%	100%	67%	67%	69%
	Specificity(TNR) --> NORMAL	73%	72%	69%	61%	71%	6%	72%	100%	67%	67%	69%
	Specificity(TNR) --> DIABETES	**85%**	**81%**	78%	84%	72%	100%	84%	0%	83%	81%	89%

ESTRATEGIA	ALGORITMO	PRECISIÓN	SENSIBILIDAD (DIABETES)
Estrategia 3. wo_PCA__UnderSampling_RUS	RF	75%	81%
	GB	74%	83%
	AB	74%	83%
	SVM-RBF	73%	84%
	MLP	72%	78%
	KNN	71%	75%
	SVM-Linear	71%	74%
	LR	71%	72%
	DT	68%	67%
	NB	52%	99%
	QDA	50%	58%

Latam Training & Solutions S.A.
Cedula Jurídica 3-101-699202
Oficentro Ejecutivo La Sabana, Torre 6 Piso 7, oficina 26-801, Sabana Sur
Teléfono: (506) 4001-3282
Email: lataminfo@linchpinpeople.com

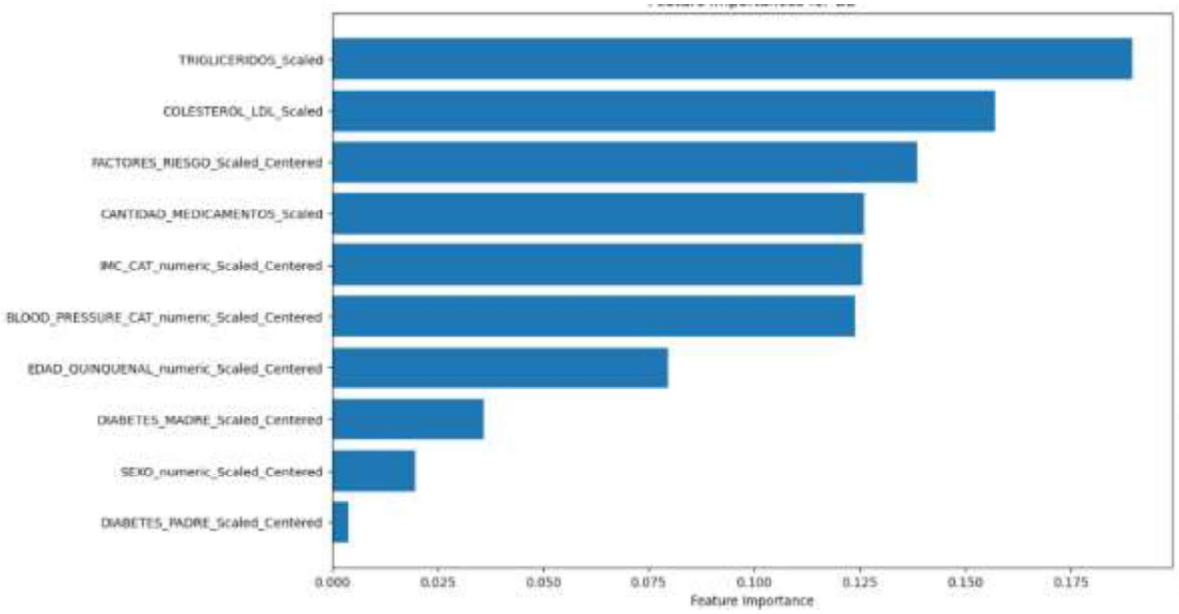
ESTRATEGIA	ALGORITMO	PRECISIÓN	SENSIBILIDAD (DIABETES)
Estrategia 4. wo_PCA_UnderSampling_NearMiss	GB	79%	73%
	SVM-RBF	79%	69%
	AB	78%	72%
	RF	77%	72%
	SVM-Linear	74%	67%
	LR	73%	69%
	MLP	73%	67%
	DT	72%	71%
	KNN	72%	61%
	NB	54%	6%
	QDA	50%	100%

Con base en los resultados anteriores, el siguiente gráfico muestra la lista de predictores ordenados por importancia para el algoritmo Gradient Boost GB:



Latam Training & Solutions S.A.
Cedula Jurídica 3-101-699202
Oficentro Ejecutivo La Sabana, Torre 6 Piso 7, oficina 26-801, Sabana Sur
Teléfono: (506) 4001-3282
Email: lataminfo@linchpinpeople.com

El siguiente gráfico muestra la lista de predictores ordenados por importancia para el algoritmo Random Forest RF:



De las anteriores graficas se puede concluir que “Colesterol LDL”, “Triglicéridos”, “Factores de Riesgo”, “IMC”, “Presión Arterial”, “Sexo”, “Diabetes Madre” y “Edad Quinquenal” tienden a tener más importancia en general. RF muestra nuevamente una distribución más balanceada de importancia que los otros algoritmos. Los predictores hereditarios “Diabetes Padre” tiene la importancia más baja en general, sin embargo, se deben utilizar en el modelo a escoger para garantizar el desempeño >70%.

Latam Training & Solutions S.A.
Cedula Jurídica 3-101-699202
Oficentro Ejecutivo La Sabana, Torre 6 Piso 7, oficina 26-801, Sabana Sur
Teléfono: (506) 4001-3282
Email: lataminfo@linchpinpeople.com

Resumen de resultados relevantes

La siguiente tabla resume los resultados más importantes y relevantes para los objetivos del proyecto. Se demuestra que es posible producir hasta 4 modelos predictores binarios (2-Clases, Normal / Diabetes) implementando equilibrio de submuestreo RUS o NearMiss, con resultados muy cercanos al utilizar todos los predictores.

Tabla No. 1 Resumen de resultados con PCA y sin PCA

Modelado sin considerar PCA

ESTRATEGIA	ALGORITMO	PRECISIÓN	SENSIBILIDAD (DIABETES)
Estrategia 3. wo_PCA__UnderSampling_RUS	RF	77%	85%
	GB	76%	86%
	SVM-RBF	75%	85%
	AB	75%	84%
	MLP	72%	78%
	SVM-Linear	72%	76%
	KNN	71%	75%
	LR	71%	72%
	DT	68%	67%
	NB	52%	99%
	QDA	50%	58%

ESTRATEGIA	ALGORITMO	PRECISIÓN	SENSIBILIDAD (DIABETES)
Estrategia 4. wo_PCA_UnderSampling_NearMiss	GB	80%	74%
	SVM-RBF	80%	72%
	RF	79%	75%
	AB	78%	74%
	SVM-Linear	74%	66%
	LR	73%	69%
	MLP	73%	67%
	DT	72%	71%
	KNN	72%	61%
	NB	54%	6%
	QDA	50%	100%

Modelado con PCA (únicamente componentes principales)

ESTRATEGIA	ALGORITMO	PRECISIÓN	SENSIBILIDAD (DIABETES)
Estrategia 3. wo_PCA__UnderSampling_RUS	RF	75%	81%
	GB	74%	83%
	AB	74%	83%
	SVM-RBF	73%	84%
	MLP	72%	78%
	KNN	71%	75%
	SVM-Linear	71%	74%
	LR	71%	72%
	DT	68%	67%
	NB	52%	99%
	QDA	50%	58%

ESTRATEGIA	ALGORITMO	PRECISIÓN	SENSIBILIDAD (DIABETES)
Estrategia 4. wo_PCA_UnderSampling_NearMiss	GB	79%	73%
	SVM-RBF	79%	69%
	AB	78%	72%
	RF	77%	72%
	SVM-Linear	74%	67%
	LR	73%	69%
	MLP	73%	67%
	DT	72%	71%
	KNN	72%	61%
	NB	54%	6%
	QDA	50%	100%

A continuación de detalle la lista de componentes principales (subconjunto de predictores) con los cuales se puede implementar el modelo para el presente proyecto:

- SEXO
- FACTORES_RIESGO

Latam Training & Solutions S.A.
Cedula Jurídica 3-101-699202
Oficentro Ejecutivo La Sabana, Torre 6 Piso 7, oficina 26-801, Sabana Sur
Teléfono: (506) 4001-3282
Email: lataminfo@linchpinpeople.com

- EDAD_QUINQUENAL
- DIABETES_MADRE
- IMC_CAT
- TRIGLICERIDOS
- BLOOD_PRESSURE
- DIABETES_PADRE
- COLESTEROL_LDL

Desde el punto de vista de modelos, el que presenta mejor desempeño en precisión y sensibilidad es el algoritmo Random Forest.

Latam Training & Solutions S.A.
Cedula Jurídica 3-101-699202
Oficentro Ejecutivo La Sabana, Torre 6 Piso 7, oficina 26-801, Sabana Sur
Teléfono: (506) 4001-3282
Email: lataminfo@linchpinpeople.com

Recomendaciones

- Implementar modelos que **contemplen** únicamente las componentes principales **PCA**, para reducir complejidad y garantizar la operacionalización del modelo.
- Si se busca un modelo con alta sensibilidad al predecir la clase “Diabetes”, el algoritmo **Random Forest** con equilibrio por submuestreo **RUS** es la opción 1 (precisión 75%, sensibilidad 81%)
- No implementar modelos predictores para diabetes tipo 2 usando los algoritmos **NB** ni **QDA** debido a que tienen métricas de precisión bajas.
- Los modelos aquí descritos están diseñados para ser implementados dentro de los sistemas de la institución (back-end), ya que los registros nuevos a predecirles las clases “Diabetes” o “Normal” deben de pasar por todo el proceso de ingeniería y estandarización de datos.

Latam Training & Solutions S.A.
Cedula Jurídica 3-101-699202
Oficentro Ejecutivo La Sabana, Torre 6 Piso 7, oficina 26-801, Sabana Sur
Teléfono: (506) 4001-3282
Email: lataminfo@linchpinpeople.com

Anexo

A continuación, se incluyen una serie de enlaces que contiene información sobre los algoritmos utilizados.

- Gradient Boosting (GB)
  https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.GradientBoostingClassifier.html
- Random Forest (RF)
  https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html
- AdaBoost (AB)
  https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.AdaBoostClassifier.html
- Support Vector Machine (SVM)
  https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html
- Linear Regression (LR)
  https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html
- K-Nearest Neighbors (KNN)
  https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.KNeighborsClassifier.html
- Decisión Tree (DT)
  https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html
- Multi-Layer Perceptron (MLP)
  https://scikit-

Latam Training & Solutions S.A.
Cedula Jurídica 3-101-699202
Oficentro Ejecutivo La Sabana, Torre 6 Piso 7, oficina 26-801, Sabana Sur
Teléfono: (506) 4001-3282
Email: lataminfo@linchpinpeople.com

learn.org/stable/modules/generated/sklearn.neural_network.MLPClassifier.html

- Naive-Bayes Gaussian (NBG)
  https://scikit-learn.org/stable/modules/generated/sklearn.naive_bayes.GaussianNB.html
- Quadratic Discrimiant Analysis (QDA)
  https://scikit-learn.org/stable/modules/generated/sklearn.discriminant_analysis.QuadraticDiscriminantAnalysis.html

Latam Training & Solutions S.A.
Cedula Jurídica 3-101-699202
Oficentro Ejecutivo La Sabana, Torre 6 Piso 7, oficina 26-801, Sabana Sur
Teléfono: (506) 4001-3282
Email: lataminfo@linchpinpeople.com