Articulo

Análisis de Riesgo de Cartera a Través de Machine Learning para Predecir la Propensión de Incumplimiento de Seguros

Carolina Hernandez Solano[1,*]

[1]Facultad de Ingeniería y Ciencias Básicas, Fundación Universitaria Los Libertadores
[*]Autor de correspondencia: chernandezs@libertadores.edu.co

LOS LIBERTADORES
FUNDACIÓN UNIVERSITARIA
Acreditación Institucional en Alta Calidad

Facultad de Ingeniería y
Ciencias Básicas

APUNTES DE
CIENCIA E INGENIERÍA

Resumen

Las aseguradoras en Colombia obtienen sus ingresos de las primas mensuales que pagan sus clientes, en este artículo se analizó una base de datos de una compañía de seguros, que ofrece un producto de seguro de vida individual con componente de ahorro; con el fin de mejorar los indicadores de riesgo de cartera, la cancelación de pólizas y aumentar los ingresos a través de un modelo para predecir la propensión de incumplimiento en el pago de las primas mensuales. Para lograr este objetivo se realizó la comparación de varios modelos teniendo como punto de referencia un modelo basado en reglas y los demás modelos se realizaron a través de la metodología machine learning, identificando el modelo Linear Discriminant Analysis como el mejor, obteniendo un resultado de recall de 0.58 % e identificando las características o variables de cada cliente que se relacionan de manera directa con el incumplimiento y con ello predecir si los nuevos clientes que tendrán incumplimiento. Con este trabajo se establece un modelo que propone a la compañía herramientas que permitan la toma de decisiones y/o definir nuevas estrategias de mercadeo.

Palabras clave: Seguros de Vida, Cartera de seguros, Machine Learning.

Recibido: 14 de octubre de 2022
Aceptado: 30 de enero de 2023
Publicado: 28 de febrero de 2023

Como citar este articulo

Hernandez-Solano, C.,"Análisis de Riesgo de Cartera a Través de Machine Learning para Predecir la Propensión de Incumplimiento de Seguros", *Revista Apuntes de Ciencia e ingeniería*, 1, 2, nov, pag 5-12. 2023. doi: 10.37511/apuntesci.v1n2a1

Copyright: ©2023 por los autores. Este artículo es de acceso abierto distribuido bajo los términos y condiciones de Creative Commons Licencia de atribución (CC BY NC SA)
(https://creativecommons.org/licenses/by-nc-sa/4.0/).

1. Introducción

Se evidencia que el sector asegurador en Colombia creció para el primer trimestre del año en un 20 % teniendo en cuenta que se emitieron primas por un total de 9.9 billones en referencia con el año 2021, de acuerdo con las cifras publicadas por el DANE al descontar el índice de precios al consumidor, se observa un crecimiento del 10.9 %, superior

al crecimiento de la economía, que para el primer trimestre de 2022 fue del 8.5 % acorde a lo publicado por la revista de (Fasecolda, 2022); además indica que para el ramo de vida individual el aumento agregado de las carteras de vivienda y de consumo fue del 15,7 % de acuerdo con las cifras de la Superintendencia Financiera al mes de febrero de 2022.

Es importante tener en cuenta el concepto de primas emitidas que hacen referencia al valor del seguro otorgado al cliente por la aseguradora por cada póliza expedida y utilidad operacional que incluye todos los ingresos de seguros y reaseguros menos los egresos de seguros y reaseguros menos los comisiones y gastos generales.

En referencia a lo anterior las aseguradoras deben mejorar la gestión del riesgo de cartera con el fin de aumentar sus ingresos, cabe resaltar que en Colombia aún no es relevante la compra de los seguros a diferencia de otros países como Estados Unidos.

En este artículo realizaremos el estudio una base de datos de una empresa aseguradora en Colombia en el ramo de vida individual con el objetivo de predecir la propensión de las personas a incumplir los pagos de las primas de seguro realizando un comparativo de las diferentes metodologías de machine learning en busca de las mejores métricas e identificar qué características de los clientes o variables influyen para que no realicen los pagos a través de un análisis descriptivo.

Con el avance tecnológico hemos evidenciado que a través de diferentes metodologías se ha realizado a través de machine learning varios modelos para predecir el riesgo en cartera en diferentes campos como el asegurador y bancario, Por ejemplo tenemos que (Gutierrez, Segovia, & Ramos, 2017) en su artículo de análisis de riesgo de caída de cartera de seguros realiza una comparación entre los modelos lineales generalizados de la estadística paramétrica y los modelos no paramétricos usando la Inteligencia Artificial y su el objetivo es demostrar la aplicabilidad del uso de la inteligencia artificial en el campo de seguros de vida, se desarrollaron diferentes modelos y logró identificar el perfil o características de los clientes que son susceptibles de realizar la cancelación o revocación de la póliza, también se demostró que esta metodología se puede usar muchas variables de entrada y no es necesario un número elevado de datos y combina fines predictivos y descriptivos.

Teniendo en cuenta lo anterior, se desarrollaron modelos machine learning y una vez identificado el mejor modelo y las variables relacionadas directamente con nuestra variable objetivo se tienen las herramientas para la creación de estrategias para reducir el riesgo en cartera identificando los posibles incumplidores para brindarles la póliza de tal manera que se ajuste al perfil, es decir por ejemplo con un menor valor de prima, mayor plazo, o menor suma asegurada.

2. Metodología

La metodología que se utiliza para desarrollar esta investigación es el siguiente: primero realizamos la depuración de la base de datos posterior a eso se realiza el análisis descriptivo. Después se procedió a realizar el análisis descriptivo donde se identificaron las variables presentan con colinealidad y por tal razón no se contemplan en el modelo y se establece que va dirigido a predecir a los clientes nuevos.

Luego generamos la división de los datos así: datos de entrenamiento "train" en un 80 % y un 20 % en los datos de prueba o "test", luego creamos un modelo Base el cual le aplicamos una metodología basada en reglas posterior a esto se realizan los modelos machine learnin con el fin de comparar los datos de test de entrenamiento de las dos metodologías a través de una matriz de confusión (Barrios, 2019) para lo cual es importante conocer las métricas que serán tenidas en cuenta:

1. Exactitud "Accuracy", que nos indica la cantidad de predicciones positivas que fueron correctas es decir verdaderos positivos y verdaderos negativos.
2. La Precisión indica el porcentaje de casos que fue correctamente clasificados de verdaderos positivos.
3. La Sensibilidad "Recall" que indica los casos positivos que el modelo ha clasificado correctamente.

Facultad de Ingeniería y Ciencias Básicas
Imagina, Diseña y Construye: Revista Apuntes de Ciencia e Ingeniería

4. la Especificidad “Specificity” que indica los casos negativos que el modelo ha clasificado correctamente.
5. El Valor-F “F1-score” que indica la capacidad explicativa de las variables sobre la variable objetivo.

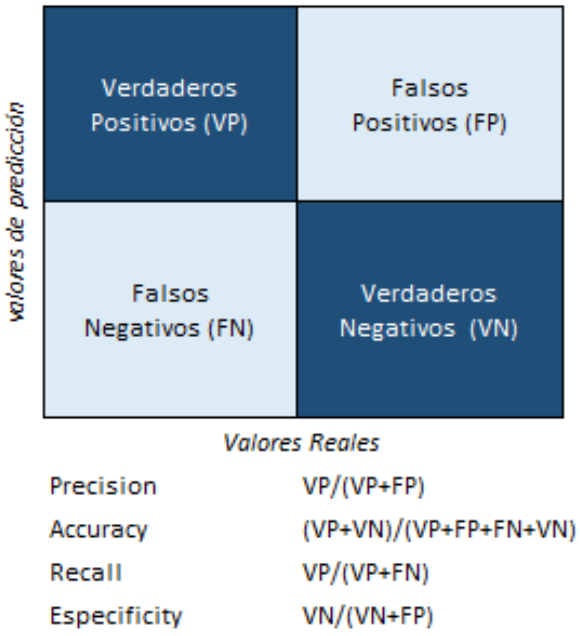

Figura 1: Matriz de confusión.

Se clasifican los clientes como incumplidores aquellos que tienen más de dos primas incumplidas y se define de esta manera porque este seguro permite tener dos primas sin pago en sus condiciones, lo anterior como beneficio teniendo en cuenta que el tiempo máximo de vigencia de la póliza es de 10 a 25 años.

Para realizar los modelos machine learning se utiliza la librería PyCaret (PyCaret Organization, 2022) la cual está diseñada para realizar la preparación de los datos así estás contengan una gran cantidad de registros, y hacer modelos y compararlos de manera automática.

Se ajustó el balance de la variable objetivo usando el algoritmo Smote (Brownlee, 2021) la cual equilibra los datos eliminando clientes que no tienen incumplimientos o creando clientes con incumplimientos.

3. Resultados

Se depuró la base de datos la cual contiene 13.034 registros y 16 variables, donde se identificaron y eliminaron los datos atípicos que corresponden a las primas con valor superior a dos millones quinientos mil pesos, para evitar sesgos en los resultados de los modelos obteniendo una base de datos de 12.719 de registros.

Como resultado de la depuración y análisis de los datos, las variables seleccionadas son las descritas en la siguiente tabla:

Facultad de Ingeniería y Ciencias Básicas
Imagina, Diseña y Construye: Revista Apuntes de Ciencia e Ingeniería

variable	Descripción
Objetivo_Ahorro	valor de póliza de seguro y objetivo de ahorro
Plazo	Tiempo de vigencia de la póliza en meses
Valor$_*Prima*$_Mensual	Valor en pesos mensual de la póliza
Incumplimientos	Se toma como "si" de dos o más incumplimientos (variable objetivo)
Edad	Edad del asegurado
Genero	Femenino o Masculino
Canal$_*de*$_ventas	Canal Agencias, Canal Empleados, Canal Intermediarios
Incremento_anual	Si o No
Ciudad	Ciudad del asegurado
Forma de Pago	RecaudoPagoVirtualInternetACHTI, TransferenciaFondos, TarjetaCredito, DebitoAutomaticoRespuesta, TrasladoGirosYPagos, TrasladoFondosSkandia, AjusteTrasladoPPA, ConsignaciónCheque ,ConsignaciónEfectivo, Tarjetadecredito,
Segmento	Clasificación de clientes por ingresos C (menor ingresos) A (ingresos medios), AA(ingresos Altos), D(Corporativos)

Tabla 1: Descripción de variables seleccionadas para el estudio.

Como resultado del análisis descriptivo se obtuvieron los siguientes datos relevantes: del total de pólizas 12719 el 18 % tienen incumplimientos con un total 2369 pólizas.

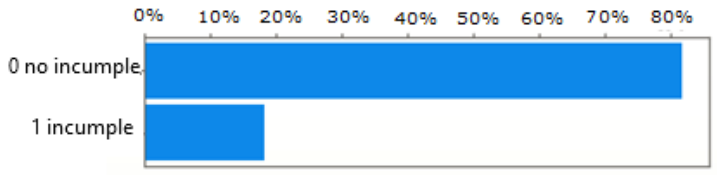

Figura 2: Porcentaje de incumplimiento.

Se evidencia que las variables que se relacionan con nuestra variable objetivo son:

Plazo, que nos indica el tiempo de vigencia o plazo de la póliza teniendo en cuenta que los clientes que escogieron una vigencia superior a los 20 años es decir 240 meses representan un 60 % de las pólizas que tienen incumplimientos.

Figura 3: Plazo vs incumplimientos.

También se evidencia que los clientes que se encuentra en los siguientes rangos de edades: de 37 a 39 años y de 50 a 55 años registran mayor porcentaje de clientes que incumplen.

Facultad de Ingeniería y Ciencias Básicas
Imagina, Diseña y Construye: Revista Apuntes de Ciencia e Ingeniería

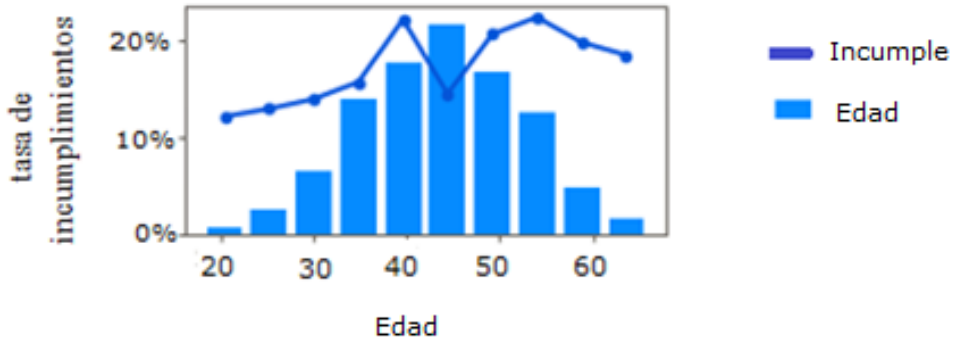

Figura 4: Edad Vs incumplimientos.

Al analizar variable valor prima mensual se evidencia que se concentra mayor incumplimiento en las primas menores de dos millones de pesos y para la variable género indica que si es hombre hombre o mujer tienen el mismo porcentaje de incumplimientos es decir que no es una variable que afecte nuestra variable objetivo.

También es importante resaltar que contamos con una variable que contiene el tiempo que llevan los clientes con su póliza es decir el tiempo transcurrido desde la emisión, que aunque no se usa en el modelo predictivo teniendo en cuenta que no aplicaría para clientes nuevos nos permite ver el comportamiento de pago y debemos resaltar que los clientes que llevan más tiempo con la póliza es decir un rango entre cinco a seis años tienen un mayor porcentaje de incumplimiento con un 30 % respecto a los demás rangos.

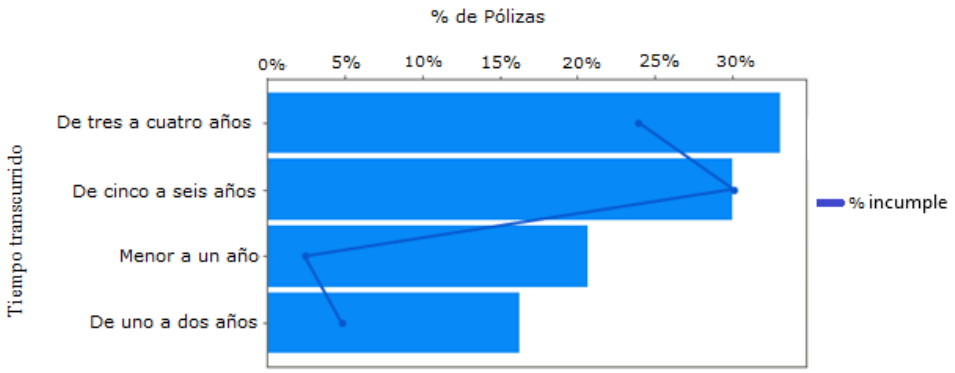

Figura 5: Tiempo transcurrido Vs incumplimientos.

3.1. Modelo basado en reglas

Se realiza un modelo basado en reglas usando la variable "Edad" y se filtra a los clientes mayores de 45 años, teniendo en cuenta que este rango registra un mayor porcentaje de clientes que incumplen acorde al análisis descriptivo.

Se selecciona la variable "Valor_Prima_Mensual" y se filtra por las primas menores a dos millones de pesos, teniendo en cuenta que allí se concentra el mayor porcentaje de incumplimiento, definidas estas variables se genera la matriz de confusión.

Facultad de Ingeniería y Ciencias Básicas
Imagina, Diseña y Construye: Revista Apuntes de Ciencia e Ingeniería

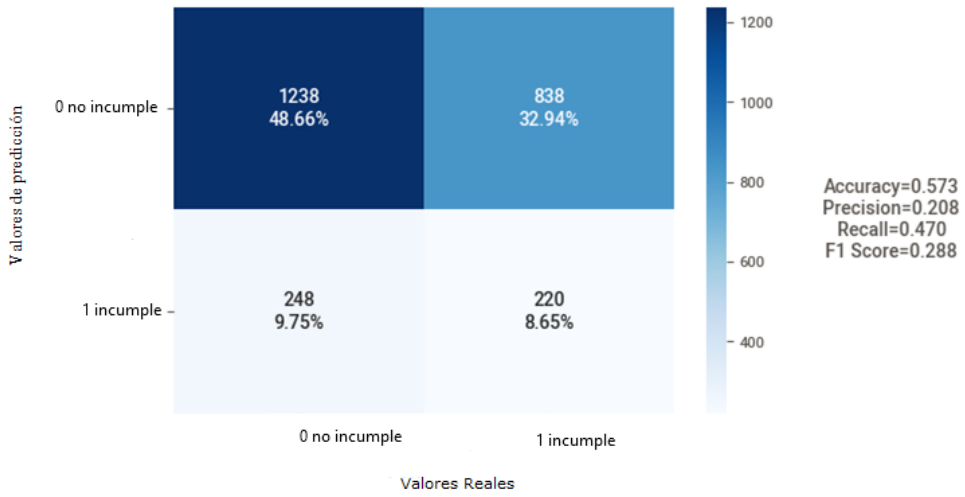

Figura 6: Matriz de confusión Modelo basado en Reglas.

Como resultado obtuvimos un F1 Score de 28 % y un Recall de 47 % lo cual nos indica que el modelo está identificado con 47 % correctamente los casos clientes que no tienen incumplimientos.

3.2. Modelos Machine Learning

Para evitar la fuga de etiqueta, solo algunas variables son utilizadas como predictoras, estas son: "Canal_de_ventas", "Ciudad", "Incremento_anual", "Forma_pago", "Segmento", "Edad", "Objetivo_Ahorro", "plazo" y "Valor_Prima_Mensual".

Una vez se obtienen los datos de entrenamiento se ejecutan los modelos y se evidencia que la base de datos se encuentra desbalanceada, es decir que tiene significativamente un número alto de clientes que no tienen incumplimientos en comparación con los que incumplen, por lo cual se ajustó el balance de la variable objetivo.

A continuación, se relaciona el resultado de los modelos y sus respectivas métricas:

Modelo	Accuracy	Recall	Prec.	F1 Score
Quadratic Discriminant Analysis	0.1893	0.9892	0.1822	0.3077
Logistic Regression	0.3339	0.8026	0.1883	0.305
Linear Discriminant Analysis	0.6195	0.6022	0.2624	0.3654
K Neighbors Classifier	0.6406	0.505	0.2545	0.3381
Naive Bayes	0.6104	0.4325	0.2144	0.285
Decision Tree Classifier	0.7513	0.3216	0.3198	0.3203
Random Forest Classifier	0.7974	0.199	0.3916	0.2629
Ada Boost Classifier	0.8037	0.1311	0.3791	0.1924
Light Gradient Boosting Machine	0.8128	0.1103	0.4467	0.1765
Basado en Reglas	0.573	0.288	0.208	0.288

Tabla 2: Comparación de Modelos.

Escogemos el modelo de Linear Discriminant Analysis teniendo en cuenta que sus métricas son equilibradas, obteniendo un Recall 60 %, f1 Score del 36 % y una precisión de del 26 % con un umbral del 50 % y realizamos la matriz de confusión con los datos de Testeo.

Facultad de Ingeniería y Ciencias Básicas
Imagina, Diseña y Construye: Revista Apuntes de Ciencia e Ingeniería

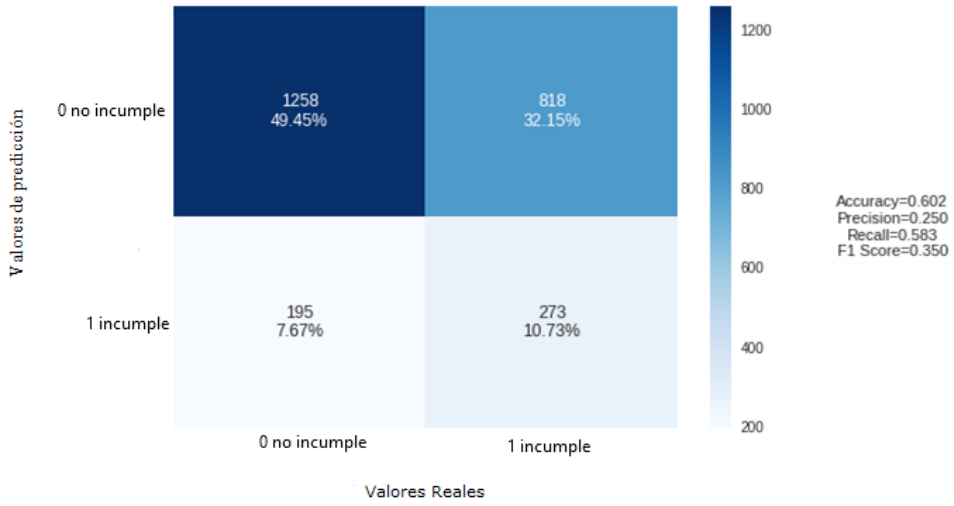

Figura 7: Matriz de confusión modelo Linear Discriminant Analysis.

Para evaluar el sobreajuste del modelo se realiza la curva de aprendizaje identificando que la curva de datos de entrenamiento no se encuentra en ningún punto con la curva de datos de validación. Sin embargo, se observa que si se tienen más datos podrían encontrarse en algún punto más adelante.

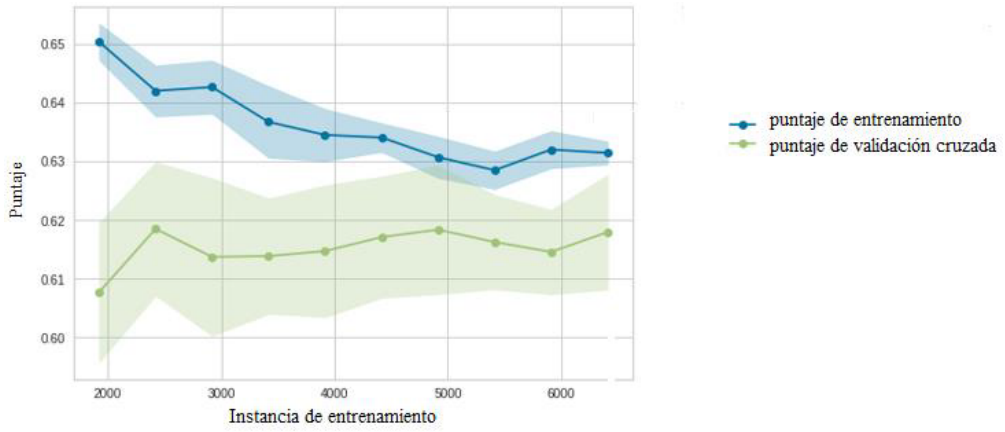

Figura 8: Curva de aprendizaje para análisis discriminante lineal.

4. Discusión de resultados

Se comparan los resultados de nuestro modelo base con el cual obtuvimos F1 Score igual a 28 % y Recall 47 % y el del modelo Linear Discriminant Analysis se obtuvo mejores métricas con F1 Score igual a 35 % y Recall 58 %.

Al consultar modelos realizados por otros autores como (Giraldo & Marin, 2021) quienes crearon un modelo para estimar el riesgo de crédito en una cartera de consumo, a través de metodologías de machine learning clasificando el modelo Random Forest con los mejores resultados, F1 Score 38 % y recall 51 %, comparando nuestro modelo se observa que se encuentra con resultados similares.

Al poder tener un modelo predictivo, la compañía puede clasificar a los prospectos de clientes como incumple y no incumple y así se disminuye el riesgo en cartera.

Para los clientes que clasifica como incumple analizar las variables como el valor de la prima, el plazo, el incremento anual y generar una nueva oferta de seguro que cambie esta clasificación y de esta manera

Facultad de Ingeniería y Ciencias Básicas
Imagina, Diseña y Construye: Revista Apuntes de Ciencia e Ingeniería

no se castigarían o rechazarían estos clientes mitigando el impacto que tendrían las ventas de seguros para no disminuir las ventas.

5. Conclusiones

Se evidencia que a través de los modelos desarrollados la metodología de machine learning arrojó un mejor resultado que el modelo basado en reglas mejorando el F1 score en 11 % y el recall en 6 % y también permite identificar las variables que se correlacionan con los incumplimientos generando así el perfil de los clientes.

De acuerdo con los resultados del mejor modelo predictivo y validando con el grupo encargado de realizar las estrategias del negocio se concluye que en la aseguradora es importante el uso de este modelo teniendo en cuenta que permite generar una caracterización de la población si incumple o no incumple a partir de las variables descriptivas y con esta información se puede crear estrategias de mercado para estos clientes.

Con el fin de optimizar y mejorar el modelo predictivo se recomienda que se incluyan más variables sociodemográficas del cliente, como por ejemplo el estado civil, número de hijos, ocupación y riesgo financiero para aumentar las métricas.

Referencias

[1] Barrios, J. (2019). La matriz de confusión y sus métricas. Online, https://www.juanbarrios.com/la-matriz-de-confusion-y-sus-metricas/.

[2] Brownlee, J. (2021). Smote for imbalanced classification with python. Online, https://machinelearningmastery.com/smote-oversampling-for-imbalanced-classification/.

[3] de Lourdes Gutiérrez Cordero, M., Segovia-Vargas, M. J., and Escamilla, M. R. (2017). Análisis del riesgo de caída de cartera en seguros: Metodologías de "inteligencia artificial" vs "modelos lineales generalizados". *Economía Informa*, 407:56–86, DOI: 10.1016/j.ecin.2017.11.004, https://doi.org/10.1016/j.ecin.2017.11.004.

[4] Durán R., V. A. and Najera A., A. A. (2022). Resultados de la industria en el primer trimestre de 2022. *Revista Fasecolda*, (186):36–44, https://revista.fasecolda.com/index.php/revfasecolda/article/view/815.

[5] Giraldo, O. (2021). Machine learning para la estimación del riesgo de crédito en una cartera de consumo. Online, https://repository.eafit.edu.co/handle/10784/29589.

[6] Gutierrez Garcia, Lizeth Daniela Trujillo Salazar, J. D. (2020). Análisis de la calidad de cartera del sistema financiero en colombia en el período 2010 -2018. Technical report, Tecnológico de Antioquia, https://dspace.tdea.edu.co/handle/tda/521.

[7] PyCaret (2022). Quickstart guide. Online, https://pycaret.gitbook.io/docs/get-started/quickstart#classification.

Facultad de Ingeniería y Ciencias Básicas
Imagina, Diseña y Construye: Revista Apuntes de Ciencia e Ingeniería