PONTIFICIA UNIVERSIDAD CATÓLICA DEL PERÚ

ESCUELA DE POSGRADO

"COMPARACIÓN DE MODELOS DE APRENDIZAJE DE MÁQUINA EN LA PREDICCIÓN DEL INCUMPLIMIENTO DE PAGO EN EL SECTOR DE LAS MICROFINANZAS"

Trabajo de Investigación para optar el grado de Magíster en Informática con mención en Ciencias de la Computación

AUTOR

Jiam Carlos López Malca

ASESOR

Cesar Augusto Olivares Poggi

JURADO

César Armando Beltran Castañon

Edwin Rafael Villanueva Talavera

LIMA – PERÚ

2,019

Comparación de modelos de aprendizaje de máquina en la predicción del incumplimiento de pago en el sector de las microfinanzas

Comparison of machine learning models in the prediction of default in the microfinance sector

Jiam López M.[1]

RESUMEN

Las instituciones financieras dedicadas a las Microfinanzas brindan sus servicios a un público objetivo que en su mayoría presentan bajos recursos económicos y/o cuyo acceso a los sistemas bancarios tradicionales es limitado, estas instituciones al desarrollarse en un contexto poco favorable los riesgos de incumplimiento en los pagos son mayores en comparación a la banca tradicional. Por tanto, se exige hacer una evaluación económica financiera con mayor grado de detalle, requiriendo para tal fin la participación de un experto del negocio que basado en información obtenida y pericia propia determine si el potencial cliente será un buen pagador. Esta forma de evaluar a un cliente ha evolucionado en el sector financiero en los últimos años, esto debido en gran medida a la aplicación de tecnologías como la inteligencia artificial y el aprendizaje de máquina, ofreciendo una singularidad que es la capacidad de aprender de los datos, demandando menos esfuerzo y participación humana, y redituando mayores niveles de precisión. Se presentan en este artículo los resultados de la experimentación realizada con los siguientes modelos de aprendizaje de maquina: Regresión Logística, XGBoost, Random Forest, Gradient Boosting, Perceptron Multicapa (MLP) y algoritmos de aprendizaje profundo para la predicción del incumplimiento de pagos, aplicándose técnicas de balanceo de submuestreo y sobremuestreo, incluida la técnica de SMOTE. Así mismo, se aplicó la técnica de One Hot Encoding para el tratamiento de variables categóricas. Los diferentes modelos de aprendizaje de maquina se aplicaron a un conjunto de datos proporcionado por una institución peruana líder en el sector de las microfinanzas, reportando los mejores resultados el modelo XGBoost, con una exactitud de 97.53% y un F1-Score de 0.1278.

Palabras clave: Aprendizaje automático, microfinanzas, microcrédito, aprendizaje supervisado.

[1] Pontificia Universidad Católica del Perú, Maestría en informática, Lima, Perú. E-mail: jiam.lopez@pucp.edu.pe

* Autor de Correspondencia

ABSTRACT

The financial institutions dedicated to Microfinance offer their services to a target audience that, for the most part, has low economic resources and/or whose access to traditional banking systems is limited, these institutions to develop in an unfavorable context the risks of non-compliance in the payments are greater compared to traditional banking, therefore it is required to make a financial economic evaluation with a greater degree of detail, requiring for this purpose the participation of a business expert that based on information obtained and own expertise determine if the potential client will be a good payer, this way of evaluating a customer has evolved in the financial sector in recent years, this largely due to the application of technologies such as artificial intelligence and machine learning, offering a uniqueness that is the ability to learn from the data, demanding less effort and human participation mana, and yielding higher levels of accuracy. This article presents the results of the experimentation carried out with the following machine learning models: Logistic Regression, XGBoost, Random Forest, Gradient Boosting, Multilayer Perceptron (MLP) and deep learning algorithms for the prediction of non-payment, applying subsampling and oversampling balancing techniques, including the SMOTE technique, and the One Hot Encoding technique was applied for the treatment of categorical variables. The different models of machine learning were applied to a data set provided by a leading Peruvian institution in the microfinance sector, with the XGBoost model reporting the best results, with an accuracy of 97.53% and an F1-Score of 0.1278.

Keywords: Machine learning, microfinance, microcredit, supervised learning.

I. INTRODUCCIÓN

Las instituciones financieras en general para un correcto desarrollo de su negocio establecen esquemas de administración y control del riesgo crediticio acorde a su propio perfil de riesgo, publico objetivo, características de los mercados en los que opera y de los productos que ofrece, siendo necesario y de cumplimiento regulatorio que cada entidad analice ciertas variables y logre determinar un perfil de riesgo que asegure la calidad de su portafolio y además permita identificar, medir, mitigar y monitorear exposiciones de riesgo y las pérdidas esperadas a fin de mantener una adecuada cobertura de provisiones y de patrimonio técnico.

Para ello, es necesario que se adopte un procedimiento de investigación y análisis, el cual permita ver reflejado en una evaluación del riesgo crediticio el conocimiento que tiene la empresa de sus clientes, para este análisis es necesario contar con suficientes datos con el fin de poder alcanzar un indicador preciso y confiable para la institución financiera y que pueda ser aplicado oportunamente en cada evaluación que realice el experto del negocio a los clientes. Esto en el corto plazo debe permitir a las instituciones financieras hacer más eficientes sus procesos de otorgamiento

de créditos, contribuyendo a mejorar indicadores como el número de clientes, recuperación de los créditos y principalmente mantener bajo el índice de morosidad en sus portafolios.

El objetivo principal de la presente investigación es identificar, a través de la comparación de diferentes algoritmos, el algoritmo más apropiado para nuestro objeto de estudio. Este algoritmo debe reportarnos los valores más altos en medidas de calidad como exactitud, precisión y exhaustividad en la predicción del incumplimiento de pagos en entidades financieras dedicadas a las microfinanzas.

II. ESTADO DEL ARTE

Varios trabajos de investigación se han hecho respecto al otorgamiento de crédito y el riesgo crediticio a través de la utilización de modelos matemáticos y aprendizaje de máquina.

Charpignon, Marie-Laure, Enguerrand Horel y Flora Tixier [8] presentan en su investigación un estudio a la creciente cantidad de empresas o nuevas empresas creadas en el campo del microcrédito, en la cual se evaluó el riesgo por defecto de sus clientes, donde el objetivo principal era predecir si un cliente experimentará incumplimiento en sus pagos (90 días a más), siendo el conjunto de datos de 100,000 clientes caracterizados por 10 variables, presentando un poder predictivo con un valor de AUC de alrededor de 85%, combinando *trees*, *bootstrap* y *gradiente boosting.* Esta investigación nos da una referencia clara del buen desempeño de técnicas de ensamble conocidas como *boosting*.

Li, Li-Hua, Chi-Tien Lin y Shin-Fu Chen [9] presentan en su investigación los resultados de aplicar una red neuronal de propagación inversa (BPN) para determinar si un cliente es buen o mal pagador. El conjunto de datos de información de préstamos pertenece a un banco comercial brasileño, el cual se utiliza para ejecutar el proceso de reconocimiento de incumplimiento crediticio, los resultados de esta investigación se muestran en la Tabla 1.

Tabla 1. Resultados de experimentación de Li, Li-Hua, Chi-Tien Lin y Shin-Fu Chen [9]

Algoritmo	*Recall Score (average)*	*F1-*
BPN (multilayer)	75.13%	57.00
Logistic Regression (LR)	50.79%	29.00

Zhu, Bing, et al [10] presentan en su investigación la aplicación del aprendizaje profundo en la calificación crediticia del consumidor a través de un modelo híbrido que combina la conocida red neuronal convolucional con el algoritmo de selección de atributos Relief [12]. Los

experimentos se llevaron a cabo en un conjunto de datos del mundo real de una empresa china de financiamiento de consumo y los resultados de esta investigación se muestran en la Tabla 2.

Tabla 2. Resultados de experimentación de Zhu, Bing, et al [10]

Algoritmo	AUC *Exactitud*	
Relief-CNN	69.89%	91.64%
Random forest	60.10%	91.40%
Logistic regresión	52.21%	85.81%

S. Islam, L. Zhou y F. Li [4] presentan en su investigación el resultado del uso de redes neuronales en el problema de *German Credit* [7], comparándolo con los valores que arroja el modelo de regresión logística. En la Tabla 3 se muestra un resumen del resultado de estos dos modelos, los cuales son de interés para el presente trabajo.

Por otro lado C. L. Huang, M. C. Chen y C. J. Wang [5], realizaron experimentación para *credit scoring* usando Support Vector Machine (SVM) como modelo de experimentación y los resultados obtenidos para el problema de *German Credit* [7] se muestran en la Tabla 4.

Tabla 3. Resultados de la experimentación de S. Islam, L. Zhou y F. Li [4]

Modelo Predictivo	Exactitud
Logistic Regression	76.4%
Neural Network	83.86%

Se debe explicar que su experimentación se basó en validación cruzada de 10 *folds* para el modelo de SVM con la finalidad de optimizar los parámetros del modelo y el conjunto de características simultáneamente. El trabajo no menciona un balance del conjunto de datos previo a la experimentación.

Tabla 4. Resultados de la experimentación de C. L. Huang, M. C. Chen y C. J. Wang [5]

Modelo + Estrategia	Exactitud
SVM + Grid search	76.00%
SVM + Grid search + F-score	77.50%
SVM + GA	77.92%

N. Ghatasheh [6] en el 2014 realizó un trabajo de experimentación con *Random Forest Trees* como modelo principal para el problema de *German Credit*. Sus trabajos incluyeron los modelos de *Random Forest*, y la comparación de resultados con otros modelos derivados del mismo como *Random Forest Adaboost* y *Bagging Random Forest*. Determina que los mejores parámetros para el número de árboles, R y M son 200, 0.3 y 0.5 respectivamente, trabajando con validación cruzada de 10 pliegues (10-folds). El diseño del experimento no menciona balanceo de los datos de entrenamiento. El resultado del trabajo de experimentación se muestra en la Tabla 5.

Tabla 5. Resultados de experimentación de N. Ghatasheh [6]

Algoritmo	Tuning	Exactitud
Random Forest	200T,R0.3,M0.5	78.4%
Random Forest	120T,15V	76.5%
Random Forest Adaboost	150T,15V	77.8%
Bagging Random Forest	100T,16V	78.4%
Default	C4.5	73.9%

T=Número de árboles
V=Variables
R=*Individual Tree Ratio*
M=Número de variables utilizadas

III. OBJETIVO Y METODOLOGÍA

En esta investigación se ha buscado identificar los posibles problemas que se pueden presentar para la predicción del incumplimiento de pagos en el sector bancario, en específico en entidades financieras dedicadas a las microfinanzas, se ha comparado entre diferentes algoritmos en vista de ver cuál es el más apropiado para nuestro objeto de estudio, así mismo, se exploró el

rendimiento de los más importantes del estado del arte.

Con respecto a la metodología que se aplicó para la presente investigación, esta consistió en cuatro etapas, estas son: exploración de datos, Pre Procesamiento, Entrenamiento y evaluación

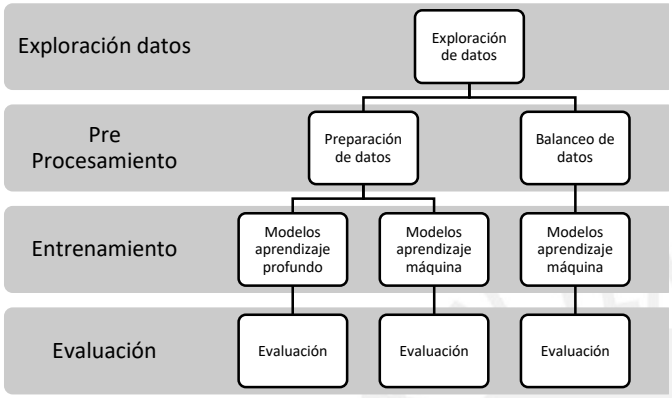

Fig. 1. Metodología de investigación

IV. EXPERIMENTACIÓN Y RESULTADOS

A. Descripción del conjunto de datos

El conjunto de datos original pertenece a una institución financiera líder en el sector de la Microfinanzas del Perú y no es de acceso público, este conjunto de datos consta de 16 atributos y 194,820 observaciones, correspondientes a préstamos desembolsados, además de un campo binario que nos indica si el préstamo cumplió o incumplió con sus pagos. Las observaciones y características se resumen de la siguiente manera:

1) Número de observaciones: 194,820.

2) Número de muestras realizadas para el presente estudio:

Entrenamiento: 146,115 observaciones.

Prueba: 48,705 observaciones.

Tabla 6. Descripción de Variables

No	Variable	Tipo	Escala	Descripción
1	Atributo_1	Variable de entrada	Cuantitativa discreta	Edad
2	Atributo_2	Variable de entrada	Cualitativa dicotómica nominal	Sexo
3	Atributo_3	Variable de entrada	Cualitativa politómica nominal	Estado Civil
4	Atributo_4	Variable de entrada	Cualitativa dicotómica nominal	Clasificación crediticia
5	Atributo_5	Variable de entrada	Cuantitativa discreta	Número de entidades sistema financiero
6	Atributo_6	Variable de entrada	Cuantitativa continua	Deuda en el sistema financiero
7	Atributo_7	Variable de entrada	Cualitativa dicotómica nominal	Tipo de crédito
8	Atributo_8	Variable de entrada	Cualitativa politómica nominal	Modulo (producto)
9	Atributo_9	Variable de entrada	Cualitativa politómica nominal	Tipo de operación (sub producto)
10	Atributo_10	Variable de entrada	Cuantitativa continua	Monto del préstamo
11	Atributo_11	Variable de entrada	Cuantitativa discreta	Plazo
12	Atributo_12	Variable de entrada	Cuantitativa continua	Total activo
13	Atributo_13	Variable de entrada	Cuantitativa continua	Total pasivo
14	Atributo_14	Variable de entrada	Cuantitativa continua	Patrimonio
15	Atributo_15	Variable de entrada	Cuantitativa continua	Ingresos
16	Atributo_16	Variable de entrada	Cuantitativa continua	Excedente
17	Atributo_17	Variable de salida	Cualitativa dicotómica nominal	Si incumplió o no

Las características contienen valores discretos y continuos, y dentro de los discretos, existen variables numéricas y categóricas. La clase de salida tiene dos valores; a saber, 0 sí no incumplió

o 1 sí incumplió. Un préstamo registra 1 como variable de salida cuando éste supera los 30 días de atraso en un periodo de tiempo de 6 meses. El conjunto de datos está claramente desbalanceado, pues existen 190,656 instancias con valor de clasificación 0 (sin incumplimiento) y 4164 con clasificación 1 (con incumplimiento), teniendo una relación de 1:45; es decir por cada 46 clientes que se les ha otorgado crédito en la entidad financiera al menos uno presenta incumplimiento en sus pagos.

B. Preparación de los datos

Uno de los primeros problemas que se va a enfrentar es preparar la data para el entrenamiento y validación de los distintos modelos de aprendizaje a utilizar. Se procede a describir la estrategia de balanceo del conjunto de datos:

1) Balanceo del conjunto de datos

En una primera instancia se evaluó balancear el conjunto de datos con la técnica de sobremuestreo, esto significaba incrementar aleatoriamente 45836 instancias correspondiente al valor 1 de la clase a predecir (si *default*) de tal manera de llegar a 50000 instancias y 190,656 instancias correspondiente al valor 0 del target (no *default*). Sin embargo, esto no hubiera sido recomendable debido a que el sobremuestreo excede al total de instancias de la clase a predecir "si default" (valor 1). Esto hubiese significado hasta más de 10 veces (uno original y otras nueve copias producto del sobremuestreo). La otra alternativa hubiese sido hacer un submuestreo de los valores de la clase a predecir, llevándolo de 190,656 a 4164 instancias, para igualar las instancias de ambas clases (4164 si *default* y 4164 no *default*). Esta estrategia hubiese significado descartar más del 98% de datos de la clasificación de "no *default*", lo cual tampoco se consideró adecuado. Por tal motivo se optó por una estrategia mixta (sobremuestreo y submuestreo) a fin de "suavizar" el impacto del balanceo.

Una técnica de sobremuestreo más eficiente seria SMOTE [11]. La ventaja de esta técnica es que permite generar nuevas observaciones (x,y) en función a la distribución de las características del conjunto de datos. En función a esta técnica se trabajó con la muestra de entrenamiento manteniendo las 142,992 instancias positivas provenientes del conjunto de datos originales y se generó 139,869 observaciones adicionales de la clase minoritaria ("si *default*") hasta completar 142,992. De esta manera el conjunto de datos quedó balanceado con 285,984 instancias (142,992 de cada clase).

Por lo tanto, la estrategia para el balanceo de los datos fue trabajada bajo la perspectiva de ambas técnicas.

C. Análisis y categorización de las características existentes

Con respecto a las variables numéricas se ha trabajado el pre procesamiento de las siguientes variables (ver Tabla 6), estas son: Deuda en el sistema financiero, Monto del préstamo, Total activo, Total pasivo, Patrimonio, Ingresos, Excedente donde se procedió con la división por 1000 y con respecto a la variable Edad, se definieron rangos, estas conversiones permitirán reducir la amplitud de dichas variables, con respecto a las demás variables numéricas se decidió mantener sus valores originales.

Para el caso de las variables categóricas se aplicó la técnica del *one-hot encoding* para transformar los datos que no son ordinales en datos numéricos binarios. Esto permite la multiplicación de las características, pasado de 16 del conjunto de datos original a 38 bajo esta técnica.

1) Algoritmos a ser utilizados

Los algoritmos o modelos seleccionados para las pruebas fueron los siguientes:

- Regresión Logística
- Redes Neuronales
- *XGBoost*
- *Random Forest*
- *Gradient Boosting*
- Redes Neuronales Profundas

D. Medidas de calidad

Se seleccionó dos medidas de calidad con el objetivo de poder comparar la validez y eficacia de los distintos modelos que se van a utilizar. El primer indicador de calidad es la exactitud de cada modelo. Se eligió este índice a pesar de que no es conveniente utilizarlo en conjuntos de datos desbalanceados, pero basados en la estrategia de balanceo que se ha planteado se decidió tenerlo como referencia y comparación entre los distintos modelos. El segundo indicador de calidad elegido es la curva ROC, el cual presenta buen funcionamiento analizando el comportamiento de clasificadores en todos los umbrales posibles, lo cual nos permitirá medir cuán óptimos son cada uno de los modelos desarrollados bajo los escenarios del conjunto de datos en su estado original (desbalanceo) y con su posterior tratamiento aplicándosele técnicas de balanceo. El tercer grupo de indicadores de calidad elegido es *Precision, Recall* y F1 principalmente por su aporte en determinar correctamente las predicciones que se realizan en un conjunto de datos altamente desbalanceado, se tiene en cuenta que el indicador de calidad F1 es la media armónica de precisión y *Recall*, no obstante, se tiene a ambos indicadores de calidad como referencia en la presente investigación.

E. Análisis de Importancia de Características

Una vez seleccionado el mejor modelo de clasificación, se realizó análisis de importancia de características, con el fin de encontrar aquellas características de mayor contribución en la predicción del Riesgo crediticio.

El modelo seleccionado para esta investigación es *Random Forest*, el cual presenta una medida de importancia de variables basada en el promedio del número de veces que se ha seleccionado cada característica en cada partición, ponderado por la mejora al cuadrado del modelo como resultado de cada partición. [2]

$$\widehat{I_j^2}(T) = \sum_{t=1}^{J-1} \widehat{\imath_t^2}\, 1(v_t = j) \qquad \textbf{(1)}$$

Donde la sumatoria es sobre los nodos no terminales t del nodo terminal J del árbol T. v_t es la variable de división asociada con el nodo t y $\widehat{\imath_t^2}$ es la correspondiente mejora empírica en error cuadrático. (1)[3].

Como una parte analítica de los datos se trabajó *features importance,* producto de este análisis y a modo referencial se presentan aquellas características que han superado el umbral del 0.1 (figura 2), valor asociado al indicador de importancia, estas serían las siguientes cuatro características (según Tabla 6): Plazo, Total activo, Patrimonio e Ingresos.

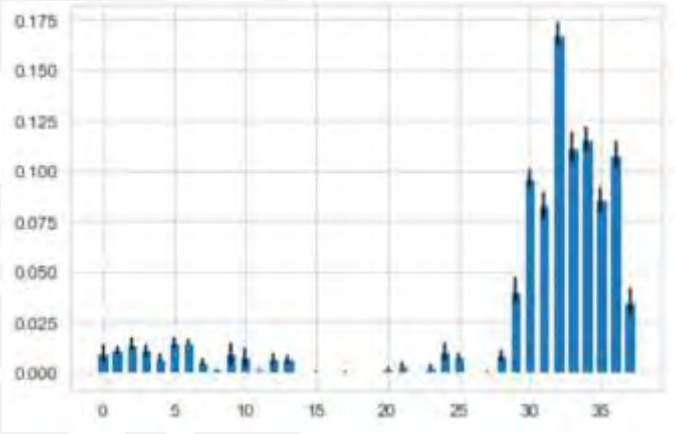

Fig. 2. Importancia de características, en el gráfico resaltamos los cuatro mejores predictores, estos son (en orden descendente): Plazo, Total activo, Patrimonio e Ingresos.

F. Resultados

Los resultados de la presente investigación se muestran a continuación estructurados por indicador de calidad para cada uno de los modelos predictivos aplicados, nuestro principal indicador de calidad es el F1-*Score* sobre indicadores como la *exactitud* presente en la mayoría de trabajos vistos como referencia en el estado del arte.

Tabla 7. Comparación de resultados de los modelos de *Machine Learning* – Indicador de calidad *Exactitud*

Modelo Predictivo	Sim. Datos Balanceado (Sub/ Sobremuestreo)	Sim. Datos Sin Balacear	Sim. Datos Balanceados (SMOTE)	Sim. Datos encoding – Sin Balancear	Sim. Datos encoding – Balanceado (SMOTE)
Regresión Logística	0.6532	0.9785	0.5485	0.9787	0.6616
Redes Neuronales	0.7238	0.9786	0.6825	0.9788	0.9043
Gradient Boosting	0.6829	0.9785	0.9739	0.9786	0.9785
Random Forest	0.9732	0.9789	0.9787	0.9787	0.9785
Xgboost	0.6814	0.9748	0.6625	0.9753	0.7300

Tabla 8. Comparación de resultados de los modelos de *Machine Learning* – Indicador de calidad *F1-Score*

Modelo Predictivo	Sim. Datos Balanceado (Sub/ Sobremuestreo)	Sim. Datos Sin Balacear	Sim. Datos Balanceados (SMOTE)	Sim. Datos encoding – Sin Balancear	Sim. Datos encoding – Balanceado (SMOTE)
Regresión Logística	0.0785	0.0000	0.0644	0.0189	0.0810
Redes Neuronales	0.0898	0.0000	0.0763	0.0337	0.1227
Gradient Boosting	0.0877	0.0206	0.1069	0.0207	0.0330
Random Forest	0.0842	0.0376	0.0814	0.0461	0.0350
Xgboost	0.0888	0.1216	0.0808	0.1278	0.0885

Tabla 9. Comparación de resultados de los modelos de *Machine Learning* – Indicador de calidad AUC

Modelo Predictivo	Sim. Datos Balanceado (Sub/ Sobremuestreo)	Sim. Datos Sin Balacear	Sim. Datos Balanceados (SMOTE)	Sim. Datos encoding – Sin Balancear	Sim. Datos encoding – Balanceado (SMOTE)
Regresión Logística	0.7397	0.6918	0.6918	0.7374	0.7477
Redes Neuronales	0.7364	0.5064	0.6979	0.7632	0.7345
Gradient Boosting	0.7651	0.7695	0.7469	0.7742	0.7456
Random Forest	0.7387	0.7337	0.7436	0.7148	0.7055
Xgboost	0.7668	0.7734	0.7438	0.7741	0.7409

Tabla 10. Comparación de resultados matriz de confusión modelos de *Machine Learning* – F1-*Score top* 4

Modelo Predictivo	*True Positive*	*False Positive*	*True Negative*	*False Negative*
Xgboost Encoding – Sin Balancear	88	248	47416	953
Xgboost Sin Balancear	85	271	47393	956
Redes Neuronales	326	3945	43719	715
Gradient Boosting	76	304	47360	965

V. DISCUSIÓN

En la experimentación el modelo con mayor exactitud fue *Random Forest* con 97.89%, en un escenario donde el conjunto de datos se encontraba sin balancear, apreciándose que el modelo fue capaz de aprender a reconocer de manera efectiva a un buen pagador, no obstante podemos apreciar en la Tabla 6 que el *F1-Score* para este modelo es uno de los más bajos que se han obtenido en la presente investigación, con respecto al AUC tenemos un valor de 73.37%, no es el valor más bajo si lo comparamos con los resultados obtenidos en los demás modelos.

Los mejores resultados fueron obtenidos con el modelo XGBoost en un escenario donde el conjunto de datos estaba sin balancear y se había aplicado la técnica *One Hot Ecoding* para las variables categóricas, consiguiendo un *F1-Score* de 0.1278, el valor más alto para este indicador en la presente investigación, con una exactitud de 97.53%, apreciándose que el valor de la exactitud no se vio penalizada como en el caso del modelo de Redes Neuronales, donde se obtuvo un *F1-Score* de 0.1227 y una exactitud de 90.43%. En la Tabla 10 podemos apreciar los modelos con los mejores indicadores, pero desde la perspectiva de una matriz de confusión, logrando el modelo de Redes Neuronales identificar un total de 326 personas con probabilidad de incumplimiento, el valor más alto, no obstante y en sintonía con la exactitud del 90.43% que obtuvo este modelo, apreciamos que se ha denegado a 3945 buenos pagadores acceso a crédito, en comparación a *XGBoost,* donde el modelo pudo identificar un total de 88 malos pagadores, el segundo valor más alto y nuevamente en sintonía con la exactitud de 97.53% que obtuvo este modelo, apreciamos que ha denegado a solo 248 (tabla 10) buenos pagadores acceso a crédito, con estos datos quedaría en la institución financiera deliberar y tomar una decisión de que modelo aplicar a sus cliente en el proceso de otorgamiento de crédito.

Con respecto al modelo de Redes Neuronales Profunda, esta se trabajó bajo la siguiente arquitectura (figura 3).

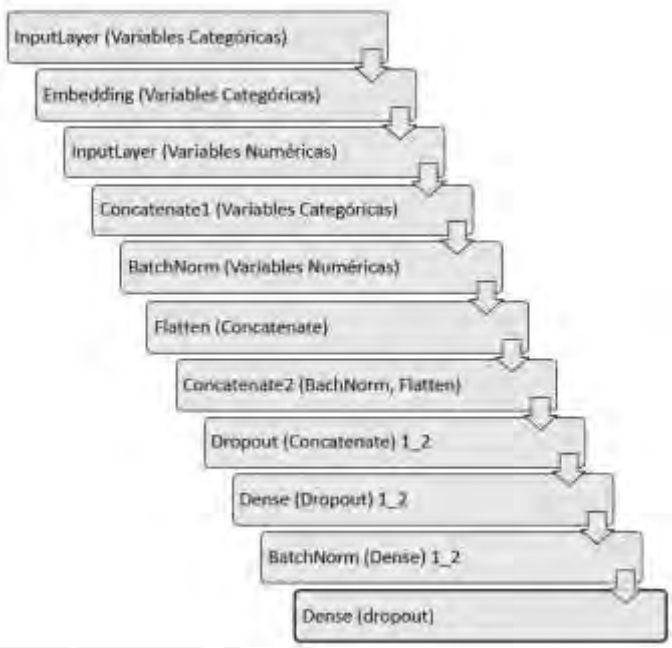

Fig. 3. Arquitectura del modelo de Red Neuronal Profunda (CNN).

El conjunto de datos empleado no estaba balanceado y se usaron *Embeddings* para las variables categóricas, en líneas generales se tuvo un buen desempeño por parte del modelo con el conjunto de datos altamente desbalanceado, este modelo se entrenó con 1000 épocas, teniendo como función de pérdida *binary_crossentropy,* distribuyéndose los pesos proporcionalmente al número de muestras de cada clase presente en el conjunto de datos con la finalidad de que el modelo no se centre en la clase mayoritaria, apreciándose en el conjunto de validación como la función de perdida marca una tendencia a la baja ante el avance de las épocas (figura 4) hasta un punto donde ya no se muestra variable, pero dentro un rango definido, entre 0.3 – 0.5, el F1-*Score* en el conjunto de entrenamiento es de 0.1238, en el conjunto de validación 9.3 y en el conjunto de pruebas 7.5, los resultados comparables serian a nivel del conjunto de pruebas, donde se puede apreciar mejores resultados con otros modelos de *Machine Learning* (ver Tabla 8).

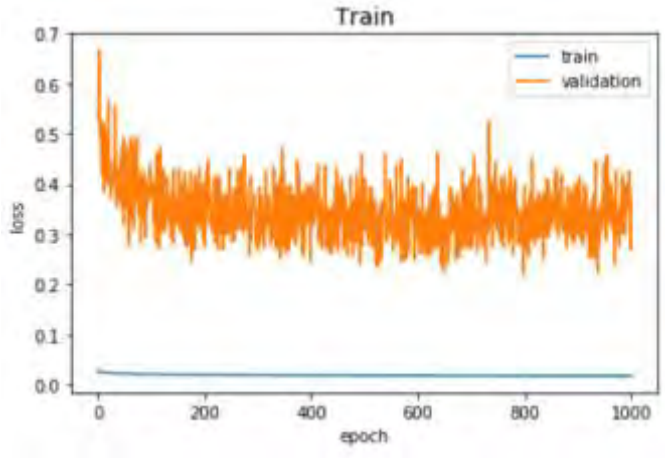

Fig. 4. Función de pérdida por *Epoch*

VI. CONCLUSIONES Y TRABAJOS FUTUROS

En presente trabajo ha tenido como objeto de estudio comparar e identificar el mejor modelo de aprendizaje de máquina para predecir el incumplimiento de pagos en el sector bancario, específicamente en instituciones financieras que se dedican a las Microfinanzas.

Random Forest ha demostrado que tiene una gran capacidad de predicción ante problemas como el que ha sido objeto de estudio en la presente investigación y por referencias que nos provee el estado del arte, no obstante, para esta investigación y en particular para el conjunto de datos que hemos utilizado, XGBoost ha logrado mejores resultados, introduciendo con esta investigación la aplicación de este modelo como referencia para abordar este tipo de problemas.

Los mejores resultados con el modelo XGBoost se dan es escenarios donde el conjunto de datos no se encuentra balanceado previamente, en nuestro caso obtuvimos 0.1278 de *F1-Score,* el valor más alto de la presente investigación, en un conjunto de datos no balanceado y con *One Hot Encoding.*

Se pudo apreciar que los algoritmos basados en árboles, como *Random Forest* o *Gradient Boosting,* el indicador *F1-Score* disminuye significativamente cuanto el conjunto de datos pasa a ser tratado con la técnica *One Hot Encoding*, pudiendo concluirse que el incremento de características influye en la *performance* de ambos algoritmos.

Revisando la comparación de las matrices de confusión a nivel de *True Positive* y *False Positive,* podemos apreciar que el modelo de redes neuronales penaliza a los buenos pagadores, por cada mal pagador que identifica (clase positiva) en relación de 1 a 12, es decir por cada mal pagador que identifica les niega a 12 buenos pagadores crédito, en comparación al modelo XGBoost que maneja una relación de solo 1 a 3.

El F1-Score de 0.1278 obtenido es un aporte de valor en nuestro caso de estudio con una muestra desbalanceada de proporciones 1:45, brindando información adicional a la que viene siendo usada en la institución financiera de origen. En efecto es mejor si lo comparamos con el F1-Score de 0.0417 que se obtiene con una predicción completamente aleatoria o el F1-Score de 0.0426 que se obtiene si se predice siempre la clase mayoritaria.

Así mismo se plantea continuar con el análisis de importancia de características y experimentar con esta primera aproximación en los diferentes escenarios y modelos que se han contemplado en la presente investigación, explorando como una técnica adicional a la propuesta, la aplicación de Redes Neuronales Convolucionales (CNN) en el análisis de importancia de características y no solo como clasificador.

En base a los resultados obtenidos se plantea seguir experimentando con el modelo de aprendizaje profundo, para lo cual se prevé incluir técnicas de análisis de importancia de características y de balanceo de datos, como parte del pre procesamiento.

REFERENCIAS

[1] N. Ghatasheh, “Business Analytics using Random Forest Trees for Credit Risk

Prediction: A Comparison Study", International Journal of Advanced Science and Technology, Vol.72, pp.19-30, Nov. 2014.

[2] J. Elith, J. R. Leathwick y T. Hastie, "A working guide to boosted regression trees", Journal of Animal Ecology, pp. 77, 2008.

[3] J. H. Friedman, "Greedy function approximation: A gradient boosting machine", Ann. Statist. 29 no. 5, pp. 1189-1232, 2001.

[4] S. Islam, L. Zhou y F. Li, Application of artificial intelligence (artificial neural network) to assess credit risk: a predictive model for credit card scoring, MSc, School of Management Blekinge Institute of Technology, 2009.

[5] C. L. Huang, M. C. Chen y C. J. Wang, "Credit scoring with a data mining approach based on support vector machines", Expert Systems with Applications, 33(4), pp. 847-856, 2007.

[6] N. Ghatasheh, "Business Analytics using Random Forest Trees for Credit Risk Prediction: A Comparison Study", International Journal of Advanced Science and Technology, Vol.72, pp.19-30, Nov. 2014.

[7] HOFMANN, D. H. (1994) Statlog (German Credit Data) Data Set. UCI Machine Learning Repository

[8] Charpignon, Marie-Laure, Enguerrand Horel y Flora Tixier. "Prediction of consumer credit risk.".

[9] Li, Li-Hua, Chi-Tien Lin y Shin-Fu Chen. "Micro-lending Default Awareness Using Artificial Neural Network.", Proceedings of the 2017 2nd International Conference on Multimedia Systems and Signal Processing. ACM, 2017.

[10] Zhu, Bing, et al. "A hybrid deep learning model for consumer credit scoring." 2018 International Conference on Artificial Intelligence and Big Data (ICAIBD). IEEE, 2018.

[11] Chawla, Nitesh V., et al. "SMOTE: synthetic minority over-sampling technique." Journal of artificial intelligence research 16 (2002): 321-357.

[12] K. Kira and L. A. Rendell, A practical approach to feature selection, Proc. 9th Int. Conf. Mach. Learn., (1992), pp. 249 – 256.