ANEXO 1

FORMATO PARA LA PRESENTACIÓN DE PROYECTOS DE INVESTIGACIÓN CON EL FINANCIAMIENTO DEL FEDU

1. Título del proyecto

EVALUACION DE LA EFICIENCIA DE MODELOS DE MACHINE LEARNING PARA OTORGAMIENTO DE CREDITOS PUNO 2022

2. Área de Investigación

Área de investigación	Línea de Investigación	Sub línea
Ingeniería y Tecnología	Sistemas, computación e informática	Inteligencia Artificial

3. Duración del proyecto (meses)

12

4. Tipo de proyecto

Individual	○
Multidisciplinario	●
Director de tesis pregrado	○

4. Datos de los integrantes del proyecto

Apellidos y Nombres	CONDORI ALEJO HENRY IVAN
Escuela Profesional	INGENIERIA DE SISTEMAS
Celular	958315508
Correo Electrónico	hcondori@unap.pe

Apellidos y Nombres	SOTOMAYOR ALZAMORA GUINA GUADALUPE
Escuela Profesional	INGENIERIA DE SISTEMAS
Celular	953620511
Correo Electrónico	guinas@gmail.com

I. **Título (El** proyecto de tesis debe llevar un título que exprese en forma sintética su contenido, haciendo referencia en lo posible, al resultado final que se pretende lograr. Máx. palabras 25**)**

EVALUACION DE LA EFICIENCIA DE MODELOS DE MACHINE LEARNING PARA OTORGAMIENTO DE CREDITOS PUNO 2022

II. **Resumen del Proyecto de Tesis (**Debe ser suficientemente informativo, presentando -igual que un trabajo científico- una descripción de los principales puntos

que se abordarán, objetivos, metodología y resultados que se esperan**)**

De acuerdo al desarrollo de las microfinanzas en el Perú, la evaluación de crediticia en microfinanzas es un procedimiento para determinar la capacidad de pago de una persona (cliente), en esta se analizan los antecedentes de los créditos otorgados en la central de riesgo del País, los estados financieros de la persona y las garantías que otorga (Chavez, 2017). De esta forma, se espera que la persona que solicita un crédito pueda pagarlo y no entrar en morosidad.
Dicho proceso en la actualidad se hace manera presencial a nivel rural del Perú, es decir, las metodologías dependen de un especialista que evalúa principalmente la capacidad de pago. Los modelos basados en metodologías presenciales han existido desde el inicio de las microfinanzas, pero con la llegada de nuevas formas de gestión de información como Machine Learning, surgen técnicas alternativas para empleando información histórica poder determinar con la misma o mayor precisión la decisión de otorgar un crédito.

Dichos algoritmos se consideran algoritmos de clasificación, que como tal tratan de agrupar de acuerdo a diversos criterios principalmente heurísticos. Por tanto algoritmos de clasificación podrían ser aplicados a la calificación crediticia, buscando contribuir a la eficiencia de las instituciones de microfinanzas, disminuyendo el riesgo crediticio. Además, se espera aplicar modelos clásicos como el análisis discriminante lineal y cuadrático, regresión logística, y metaheurísticas como redes neuronales, máquinas de soporte vectorial, árboles de clasificación, entre otros (Cubiles-de-la-vega, Blanco-oliver, Pino-mejías, & Lara-rubio, 2013). Este proceso de emplear algoritmos de machine learning se ha ido incrementado en todas las áreas y también en evaluación crediticia, por lo que existe un conjunto varios de algoritmos que se puede aplicar. Por lo tanto el presente estudio plantea realizar una revisión de la literatura sobre algoritmos de clasificación basados en machine learning y determinar su efectividad a través de su nivel de asertividad al análisis de riesgo al otorgar un crédito en instituciones financieras, para poder determinar su aplicabilidad desde el punto de vista teórico a las microfinanzas, referidas al otorgamiento de créditos pecuarios, a fin de mostrar los más asertivos para el contexto en estudio.

III. Palabras claves (Keywords) (Colocadas en orden de importancia. Máx. palabras: cinco)

Microcréditos rurales, algoritmos de clasificación machine learning, asertividad, métricas de evaluación.

IV. Justificación del proyecto (Describa el problema y su relevancia como objeto de investigación. Es importante una clara definición y delimitación del problema que abordará la investigación, ya que temas cuya definición es difusa o amplísima son difíciles de evaluar y desarrollar)

La decisión de otorgar o no un crédito, es determinante para garantizar el correspondiente pago futuro del dinero otorgado, a fin de evitar índices de morosidad. Por lo que considera una tarea compleja de realizar, por esta razón, muchas empresas optan por la construcción de modelos, los cuales permitan identificar si un determinado microcrédito puede ser otorgado o no (Tesén, 2017) y más aún en el sector microfinanciero rural. Por lo que, se considera que es

posible mejorar la eficiencia del proceso de otorgamiento, a través del empleo de algoritmos de clasificación basado en Machine learning, que se puede determinar a través de poner a prueba dichos algoritmos con diversas métricas, y determinar su nivel de asertividad para el presente escenario, lo que mejoraría el desempeño de las entidades que necesitan determinar con la mayor precisión posible la decisión de otorgar un crédito o no, y a su vez mejorar las cifras de inclusión financiera en el sector rural peruano.

V. Antecedentes del proyecto (Incluya el estado actual del conocimiento en el ámbito nacional e internacional. La revisión bibliográfica debe incluir en lo posible artículos científicos actuales, para evidenciar el conocimiento existente y el aporte de la Tesis propuesta. Esto es importante para el futuro artículo que resultará como producto de este trabajo)

En esta sección se mencionan algunos algoritmos y/o modelos usados en Machine Learning.

Modelos de *Machine Learning*

1.1 Regresión Logística

Según Ng (2018) la regresión logística es un algoritmo de clasificación de aprendizaje supervisado en la que establece una relación de variables independientes representada por X y una variable dependiente conocida como y a través de la siguiente ecuación:

$$z^{(i)} = w^T x^{(i)} + b$$

$$a^{(i)} = \sigma\left(z^{(i)}\right)$$

$$L\left(a^{(i)}, y^{(i)}\right) = -y^{(i)} log\left(a^{(i)}\right) - \left(1 - y^{(i)}\right) log\left(1 - a^{(i)}\right)$$

$$J = \frac{1}{m} \sum_{i=1}^{m} L\left(a^{(i)}, y^{(i)}\right)$$

Donde:

$$x^{(i)}: Iésimoejemplodeentrenamiento.$$

$$y^{(i)}: Etiquetadesalidaeliésimoejemplodeentrenamiento.$$

$$W: Elpesovectorial.$$

$$b: Bias, variablequeapoyaengeneralizacióndelmodelo$$

$$a^{(i)}: Predicción$$

$$\sigma: Funciónsigmoide$$

$$L: Funciónerror$$

$$J: Funcióncosto$$

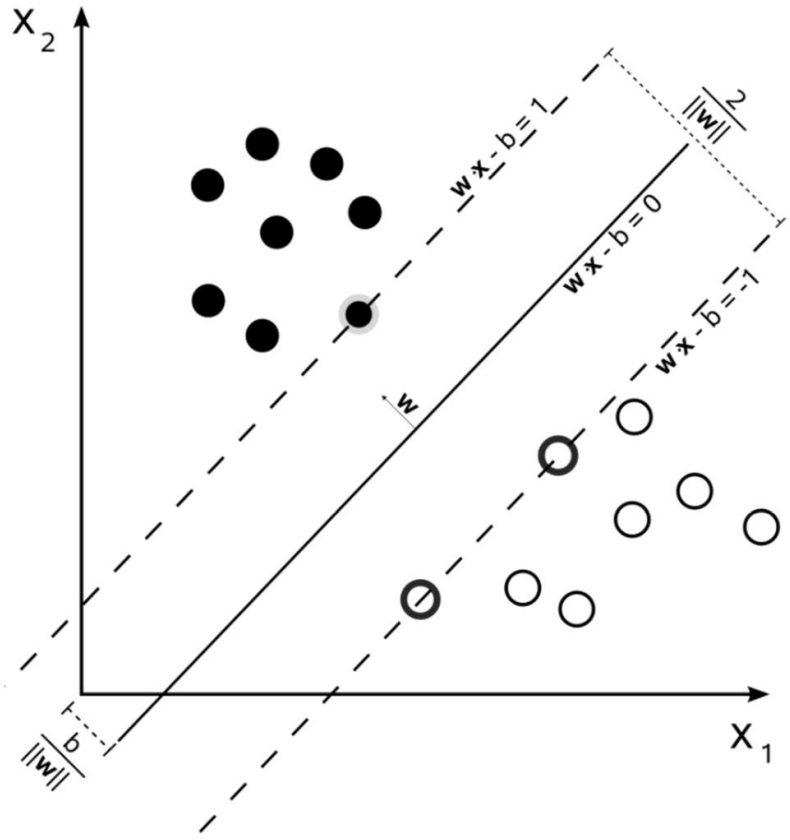
m: $Número de ejemplos de entrenamiento en el dataset$

La regresión logística es uno de los modelos de clasificación más conocidos y utilizados en análisis de riesgo crediticio, como los trabajos de Millán & Caicedo (2018), Addo et al. (2018), Kalayci et al. (2018), Arango & Restrepo (2017), Valencia (2017), Kruppa et al. (2013) entre otros.

1.2 Support Vector Machine (SVM)

Support Vector Machine es un algoritmo de aprendizaje supervisado, identificado como clasificador binario, lo que implica que las etiquetas de clasificación deben ser 0 o 1, Verdadero o Falso, Azul o Verde, entre otros, este algoritmo utiliza un enfoque diferente al probabilístico que es usado en otros algoritmos de *Machine Learning*, permitiendo razonar de una forma geométrica basándose en productos internos y proyecciones (Deisenroth et al., 2019). A su vez indican que es un poderoso método de aprendizaje automático en la clasificación de datos (Liang et al., 2016). En general, la idea principal de *SVM* es determinar el hiperplano de separación que maximiza el margen entre dos clases de datos de entrenamiento, según la teoría de la optimización, dicho hiperplano óptimo se especifica mediante el vector de peso *w* y el sesgo *b* (Figura 4), que son soluciones del problema de optimización restringida (Nguyen, 2016).

Figura *1*: Clasificación a través de Support Vector Machine



Fuente: Nguyen (2016)

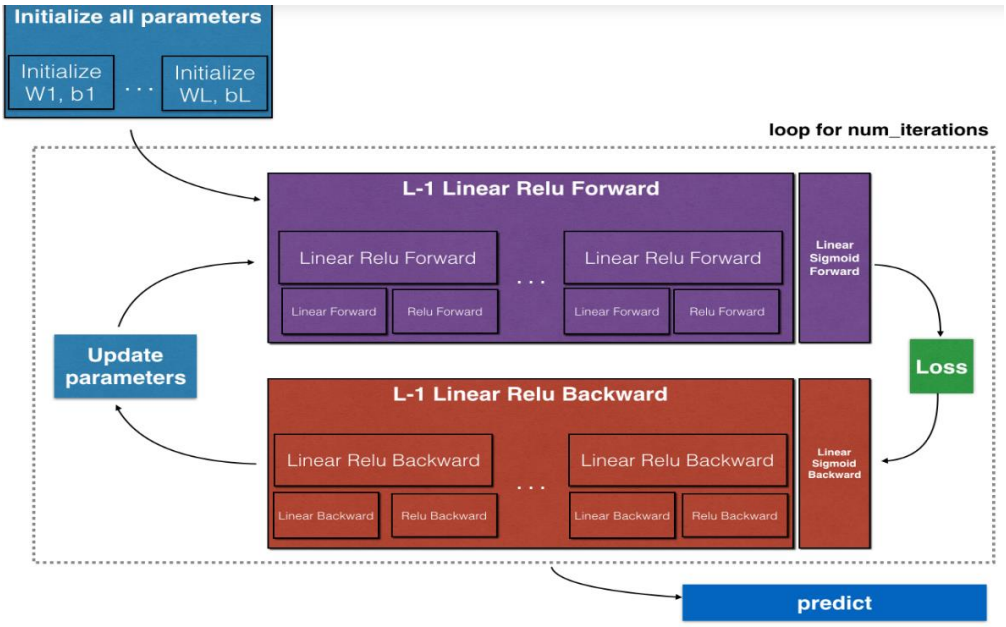
SVM fue utilizado en diversas investigaciones relacionadas al tema de riesgo crediticio tales como: Chiranjit & Andreas (2017), Kalayci et al. (2018), Flores & Ramon (2014), Turkson et al. (2016) entre otros.

1.3 Artificial Neural Network (ANN)

ANN es un modelo computacional inspirado en la biología, consiste en elementos de procesamiento (llamados neuronas) y conexiones entre ellos con coeficientes (pesos) unidos a las conexiones. Estas conexiones constituyen la estructura neuronal y se unen a esta estructura los algoritmos de entrenamiento y recuperación. *ANN* se denominan modelos conexionistas debido a las conexiones que se encuentran entre las neuronas (Shanmuganathan, 2016).

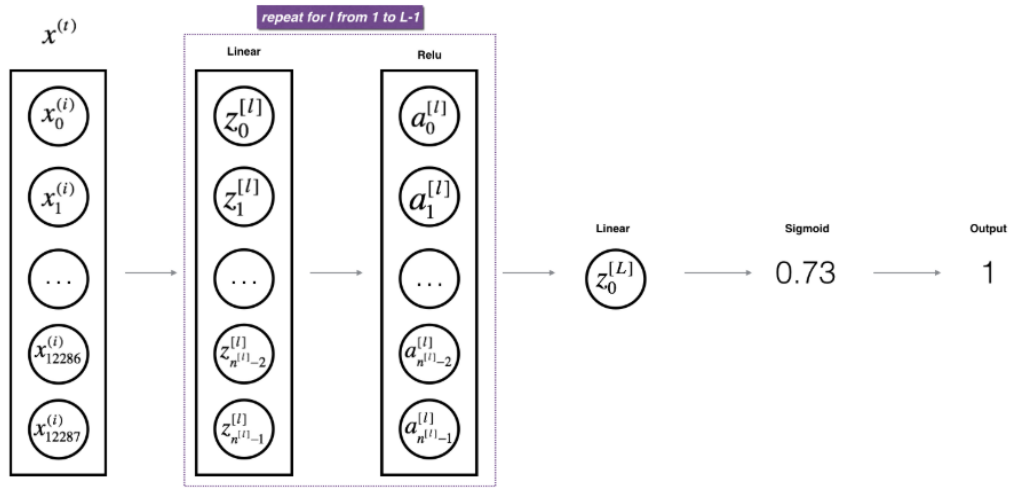
Ng (2018) presenta el modelo integral de una red neuronal artificial donde se integra todos los componentes de dicho modelo como la inicialización de variables, el algoritmo de *Forward Propagation*, la función de pérdida, el algoritmo de *Back Propagation* y la actualización de los parámetros.

Figura *2*: Modelo integral de una ANN



Fuente: Ng (2018).

Figura 3: Algoritmo Forward Propagation en una ANN



Fuente: Ng (2018).

La Figura 6 muestra el algoritmo *Forward Propagation*, donde matemáticamente se expresa de la siguiente forma:

$$z^{[1](i)} = W^{[1]}x^{(i)} + b^{[1]}$$

$$a^{[1](i)} = f\left(z^{[1](i)}\right)$$

$$z^{[2](i)} = W^{[2]}x^{(i)} + b^{[2]}$$

$$a^{[L-1](i)} = sigmoid\left(z^{[L-1](i)}\right)$$

$$y^{(i)}_{prediction} = \begin{cases} 1, si a^{[L-1](i)} > 0.5 \\ 0 \end{cases}$$

Donde:

$L: Número de capas en la neuronal.$

$x^{(i)}: Iésimo ejemplo de entrenamiento.$

$W^{[l]}: Matriz de pesos en la capa l.$

$b^{[l]}: Bias en la capa l, variable que apoya en generalización del modelo.$

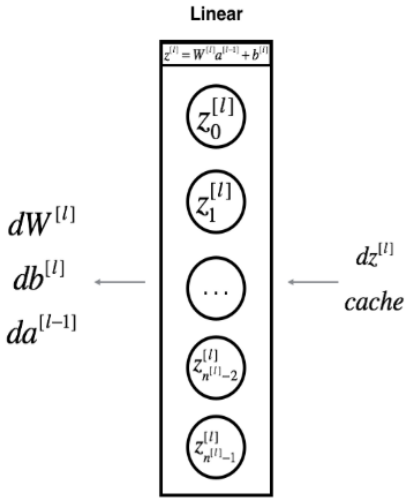
$a^{[l]}: Predicción en la capa l.$

$z^{[l]}: Resultado de la función lineal en la capa l.$

$f: Función de activación$

En la Figura 7 muestra la representación gráfica del algoritmo *Backward Propagation.*

Figura 4: Algoritmo Backward Propagation en una ANN



Fuente: Ng (2018).

La expresión matemática de *Backward Propagation* se expresa a través de la siguiente forma:

$$dW^{[l]} = \frac{\partial L}{\partial W^{[l]}} = \frac{1}{m} dZ^{[l]} A^{[l-1]T}$$

$$db^{[l]} = \frac{\partial L}{\partial b^{[l]}} = \frac{1}{m} \sum_{i=1}^{m} dZ^{[l](i)}$$

$$dA^{[l-1]} = \frac{\partial L}{\partial W A^{[l-1]}} = W^{[l]T} dZ^{[l]}$$

Donde:

$$l: Númrerodecapadelaneuronal$$

$$W^{[l]}: Matrizdepesosenlacapal.$$

$b^{[l]}: Biasenlacapal.$

$A^{[l]}: Predicciónenlacapal.$

$Z^{[l]}: Resultadodelafunciónlinealenlacapal.$

$L: Funciónerror.$

$$m: Númerodeejemplosdeentrenamientoeneldataset.$$

1.4 Decision Tree

Según Shalev & Shai (2014) definen *Decision Tree (DT)* como un método de aprendizaje supervisado no paramétrico utilizado para la clasificación y la regresión. El objetivo es crear un modelo que prediga el valor de una variable objetivo mediante el aprendizaje de reglas de decisión simples inferidas de las

características de los datos.

Los criterios de decisión de un árbol de decisión se dan por la siguiente expresión:

$$\Delta I(t) = I(t) - \frac{N_{tS}}{N_t} I(t_S) - \frac{N_{tN}}{N_t} (t_N)$$

Donde:

$$t: representaunnodoenelárbol$$

$$N_t: Númerototaldemuestrasenelnodeopadret.$$

$$N_{tS}: Númerototaldemuestrasenviadosalnodo SI.$$

$$N_{tN}: Númerototaldemuestrasenviadosalnodo No.$$

1.5 k-Nearest Neighbor (kNN)

Según Ng (2018) menciona que *kNN* es uno de los muchos algoritmos de aprendizaje supervisado utilizados en el campo de la minería de datos y el aprendizaje automático, es un clasificador donde el aprendizaje se basa en cuán similar es un dato a otro. El entrenamiento está formado por vectores de *n* dimensiones.

Para calcular la distancia entre dos puntos, la nueva muestra y todos los demás datos que tiene en su conjunto de datos existen varias formas de obtener este valor, donde la distancia euclidiana ya que es uno de los más usados, y está determinado por la siguiente expresión:

$$\rho(x, x') = \sqrt{\sum_{i=1}^{n}}$$

Donde:

$$\rho: funcióndedistanciaeuclidiana.$$

$x_i: i - ésimoejemplodeentrenamiento.$

$x'_i: i - ésimoejemplodepredicción.$

$$n: cantidaddeatributos$$

1.6 Random Forest

Breiman (2001) propuso el algoritmo de bosque aleatorio como un método de clasificación y regresión de propósito general. Los bosques aleatorios son una

combinación de predictores de árboles, de modo que cada árbol depende de los valores de un vector aleatorio muestreado de forma independiente y con la misma distribución para todos los árboles en el bosque.

Formalmente Breinman (2001) define un bosque aleatorio como un clasificador que consiste en una colección de clasificadores estructurados en árbol:

$$\{h(x, \Theta_k)\}$$

Donde:

$$h: Clasificadorestructuradoenárbol$$

$$k: k-ésimoárbol$$

$$x: Vectordeatributosdeentrada$$

$$\Theta_k: Vectoresaleatoriosindependientesdelárbolk.$$

Ayma (2019) indica que la salida final, de un bosque aleatorio, corresponde a la clase de moda entre los árboles, cada árbol del bosque es diferente respecto a los atributos y conjunto de entrenamiento, en donde la selecciona aleatoria de atributos de un espacio, $X \epsilon R^l$, se crean t árboles de decisión con diferentes espacios de atributos $X_s \epsilon R^{ls}, R^{ls} \subseteq R^l$.Así mismo indica que la selección aleatoria de muestras de entrenamiento, X_e, consiste en dividir X_e en T subconjuntos X_t, para entrenar cada árbol de decisión t con un subconjunto de entrenamiento X_t.

En los árboles estándar, cada nodo se divide utilizando la mejor división entre todas las variables. En un bosque aleatorio, cada nodo se divide utilizando el mejor entre un subconjunto de predictores elegidos aleatoriamente en ese nodo. Esta estrategia se desempeña muy bien en comparación con muchos otros clasificadores, incluido el análisis discriminante, las máquinas de vectores de soporte y las redes neuronales, y es robusta contra el sobreajuste (Liaw & Wiener, 2002).

VI. Hipótesis del trabajo (Es el aporte proyectado de la investigación en la solución del problema)

Es posible determinar los modelos de machine learning de clasificación con mejor ajuste de eficiencia para determinar el otorgamiento de créditos.

VII. Objetivo general

Evaluar la eficiencia de modelos de machine learning para otorgamiento de créditos Puno 2022

VIII. Objetivos específicos

Revisar los modelos de machine learning clasificadores.
Determinar los criterios de medición.
Establecer la evaluación de los modelos eficientes para la evaluación de microcréditos.

IX. Metodología de investigación (Describir el(los) método(s) científico(s) que se empleará(n) para alcanzar los objetivos específicos, en forma coherente a la hipótesis de la investigación. Sustentar, con base bibliográfica, la pertinencia del(los) método(s) en términos de la representatividad de la muestra y de los resultados que se esperan alcanzar. Incluir los análisis estadísticos a utilizar)

La metodología a emplear se basa en la revisión y el análisis de los modelos teóricos con fuerte respaldo científico, para luego en base a criterios propuestas determinar su caracterización aplicable a créditos rurales pecuarios, siendo exploratorio

Enfoque:
Evaluación cualitativa, cuantitativa.
Periodo:
Tres últimos años.
Alcance:
Los objetivos del presente estudio son descriptivo exploratorio.

Diseño de Investigación

Es de tipo estudio de casos e investigación de literatura; pues trata de determinar características de los modelos que pueden ser aplicables a un caso.

Técnicas de Investigación

Entre las principales técnicas empleadas en el presente trabajo de investigación, se considera la encuesta, que permitirá recabar la información de la percepción de los usuarios.
Además la observación directa y entrevista para el monitoreo del proceso de desarrollo del experimento.
Revisión bibliográfica

Instrumentos de Recolección de Datos
Para la presente investigación, se empleará el cuestionario, que será desarrollado a partir de las variables e indicadores de la operacionalización de variables de la información histórica de créditos.

Procesamiento y Análisis de Datos

Para el procesamiento, se utilizará algún software especializado, como Phyton y R, a fin de registrar las encuestas aplicadas, bajo un formato predefinido de captura de información en base al cuestionario formulado.

X. Referencias (Listar las citas bibliográficas con el estilo adecuado a su especialidad)

Aceituno, M., 2019. Modelo predictivo de analisis de riesgo crediticio usando Machine Learning en una entidad del sector microfinanciero. ´ Ph.D. thesis. Universidad Nacional del Altiplano.

Addo, P., Guegan, D., Hassan, B., 2018. Credit risk analysis using machine and deep learning models. SSRN doi:10.2139/ssrn.3155047.

Albon, C., 2018. Machine Learning with Python Cookbook. O'Reilly Media, Inc.
Chakraborty, C., Joseph, A., 2017. Machine learning at central banks. Technical Report. Bank of England.

Condori-Alejo, H.I., Aceituno-Rojo, M.R., Alzamora, G.S., 2021. Rural micro credit assessment using machine learning in a peruvian microfinance institution. Procedia Computer Science 187, 408–413. doi:https://doi.org/10.1016/j.procs.2021.04.117.

Flores, R., Ramon, J., 2014. Modelling credit risk with scarce default data: on the suitability of cooperative bootstrapped strategies for small low-default portfolios, in: Journal of the Operational Research Society.

Guajardo, J., 1991. Estrategias y tecnicas para optimizar el crédito y la cobranza. Master's thesis. Universidad Autónoma de Nuevo León.

Hossin, M., M.N, S., 2015. A review on evaluation metrics for data classification evaluations. International Journal of Data Mining & Knowledge Management Process 5, 01–11. doi:10.5121/ijdkp.2015.5201.

Morales, J., Morales, A., 2014. Credito y cobranza. Grupo Editorial Patria.
Mueller, A., Guido, S., 2016. Introduction to Machine Learning with Python. O'Reilly Media, Inc.

Tesen, A., 2017. Eficacia de los modelos de aprendizaje de máquina para evaluar el riesgo crediticio de personas naturales en una institución financiera de Chiclayo. Ph.D. thesis. Universidad Nacional de Santa.

XI. Uso de los resultados y contribuciones del proyecto (Señalar el posible uso de los resultados y la contribución de los mismos)

Permitirá contar una definición de eficiencia de cada modelo de clasificación debidamente caracterizados en cuanto a la aplicabilidad en el otorgamiento de créditos en el sector rural.

XII. Impactos esperados

i. Impactos en Ciencia y Tecnología

Plantea una propuesta teórica de los modelos aplicables a créditos pecuarios en el sector rural basado en machine learning.

ii. Impactos económicos

La evaluación de modelos permitirá determinar el modelo de machine learning aplicable a la evaluación de créditos pecuarios en el sector rural con la correspondiente reducción de costos.

iii. Impactos sociales

Mejorar el proceso de inclusión financiera en el área rural y acceso al crédito de manera más ágil.

iv. Impactos ambientales

No genera impacto o costo ambiental

XIII. Recursos necesarios (Infraestructura, equipos y principales tecnologías en uso relacionadas con la temática del proyecto, señale medios y recursos para realizar el proyecto)

Centro de experimentación, con una entidad financiera que realice el otorgamiento de créditos en el área rural
Phyton y R, R Studio.
Tecnologías de encuestas electrónicas.
Computadores con capacidad de procesamiento de volumen de información.

XIV. Localización del proyecto (indicar donde se llevará a cabo el proyecto)

Puno

XV. Cronograma de actividades

ACTIVIDADES	CALENDARIO 2022											
	1er mes	2do mes	3er mes	4to mes	5to mes	6to mes	7mo mes	8vo mes	9no mes	10mo mes	11vo mes	12vo mes
Selección de antecedentes Documentarios	X	X										
Identificación del Problema y Formulación de los Objetivos			X									
Antecedentes Bibliográficos			X	X	X							
Análisis de los modelos			X	X								
Caracterización general del área de microcréditos			X									
Revisión, caracterización del cada modelo de ML y ejecución de bases				X	X	X	X	X	X	X		
Evaluación de modelos										X	X	

Desarrollo del Informe											X	X
Presentación												X

XVI. Presupuesto

OBJETIVOS

ETAPAS DE PROYECTO	DIAS	HORAS	IMPORTE S/
Análisis Descriptivo	33	100	2000
Construcción	20	160	2200
Implementación y ejecución de bases	30	300	3000
Pruebas	45	150	900
Verificación	30	150	950
Sub - Total	***158***	***860***	***9050***

MATERIALES

DESCRIPCION DE MATEARIALES	Unid	Cantidad	IMPORTE S/
Bibliografía e Información	Unid	2	150
almacenamiento	unid	1	17,5
Papel Bond 80 Gr.	millar	10	250
Plumones de Pizarra	unid	5	17,5
Lapiceros	unid	3	4,5
Agenda A4	unid	1	10
Imprevistos	%	10	44,95
Sub - Total			***410***

SERVICIOS

DESCRIPCION DE SERVICIOS	Unid	Cantidad	IMPORTE S/
Internet	mes	100	1200
Capacitación	Unid	1	350
Digitación	hojas	500	400
Impresión	hojas	500	600
Empastado y Anillados	Unid	5	300
Imprevistos	%	5	513
Sub - Total			***3163***

OTROS

DESCRIPCION	Unid	Cantidad	IMPORTE S/
Investigación	Unid	1	300
Viáticos (lugares de investigación)	viajes	2	500
Movilidad (lugares de investigación)	viajes	2	100
Imprevistos	%	5	90
Sub – Total			***990***

RESUMEN

OBJETIVOS	*9050*
MATERIALES	*410*
SERVICIOS	*3163*

OTROS	990
TOTAL	**12722**