UNIVERSIDAD ESAN

FACULTAD DE INGENIERÍA

INGENIERÍA DE TECNOLOGÍAS DE INFORMACIÓN Y SISTEMAS

Predicción del estado de financiamiento de proyectos de tecnología en web de crowdfunding Kickstarter mediante modelo(s) de Aprendizaje Automático

Tesis para optar el Título de Ingeniero de Tecnologías de Información y Sistemas que presenta:

Alonso Augusto Puente Ríos

Asesor: Marks Arturo Calderón Niquin

Lima, 27 de junio de 2020

Esta tesis denominada:

PREDICCIÓN DEL ESTADO DE FINANCIAMIENTO DE PROYECTOS DE TECNOLOGÍA EN WEB DE CROWDFUNDING KICKSTARTER MEDIANTE MODELO(S) DE APRENDIZAJE AUTOMÁTICO

ha sido aprobada.

. . . . . .

(Jurado Presidente)

. . . . . .

(Jurado)

. . . . . .

(Jurado)

Universidad ESAN

2020

PREDICCIÓN DEL ESTADO DE FINANCIAMIENTO DE PROYECTOS DE TECNOLOGÍA EN WEB DE CROWDFUNDING KICKSTARTER MEDIANTE MODELO(S) DE APRENDIZAJE AUTOMÁTICO

Agradecimiento y dedicatoria

Durante la inducción en la empresa en la cual realicé mis segundas prácticas pre-profesionales, se realizaron varias actividades, entre ellas, una que me marcó positivamente. Esta consistía en comparar los tiempos de llegada de un punto a otro de una persona corriendo. Se caracterizó porque quien asumió el reto tuvo presente en su mente las personas y las razones por las cuales todos los días lucha y son su principal fuente de motivación.

Por ello, quiero dedicar este gran esfuerzo personal de trabajo de tesis a quienes siempre han estado a mi lado en los mejores y peores momentos, aquellos críticos en que definen el destino. Mi amada hermana Clarisabel, mis queridos padres Augusto e Isabel, mi familia en especial mis abuelos; y mis pocos, pero verdaderos y leales amigos de la universidad, colegio y trabajo. Todos ellos han sido y son cada uno, piedra fundamental en el desarrollo de mi ser como persona y profesional, así como también seres con los cuales siempre comparto gratos momentos. Su presencia en mi vida no ha sido una suerte más sino parte de mi destino. Asimismo, luchar por mis sueños y mi país, y pensar cada día en solidificar su planificación me motivan emocionalmente hasta en aquellos momentos en que parece haber imposibles.

Quiero concluir esta sección, muy especial para mí, agradeciendo también a mi alma máter, la Universidad Esan, y al Programa Nacional de Becas (Pronabec) por hacer que estos 5 años entre el 2015 y 2019 sean mágicos y muy fructíferos. Tuve la oportunidad no solo de incrementar y potenciar mis conocimientos en distintas áreas académicas sino también de aprender de excelentes profesionales como mis profesores, conocer grandes amigos dentro y fuera de su campus (desde el primer ciclo como cachimbo hasta el último ciclo, en el CADE Universitario 2019, estudiantes de diferentes universidades y otras partes del Perú), ponerme a prueba en el exterior (en el II Congreso Internacional de Investigación en Colombia) y formar parte de la gran familia UE.

Por todos ellos, simplemente gracias.

Índice general

Índice de Figuras

Índice de Tablas

Resumen

Conocer el destino del financiamiento de proyectos siempre ha sido el principal deseo de todos los emprendedores que los promocionan en Internet, en especial, de la categoría de tecnología por ser los que presentan las ratios más bajas de éxito debido a sus altas metas que buscan alcanzar. El presente trabajo de investigación se basó en construir un modelo predictivo cuyo objetivo es la de estimar el éxito o fracaso de financiamiento de proyectos tecnológicos en la plataforma de crowdfunding Kickstarter durante la duración de su campaña a partir de su metainformación, imagen y/o descripción. Para ello, se crearon modelos de Aprendizaje Automático (SVM, MLP y CNN) para cada una de estas partes. Luego de analizar todos los modelos con las mismas métricas mencionadas en el décimo antecedente, se concluyó que solo los de metainformación tuvieron niveles excelentes de acuerdo a sus puntajes AUC (0.8377 para SVM y 0.7043 para MLP), mientras que los modelos de descripciones tuvieron un rendimiento regular (0.6746 para SVM con TF-IDF y 0.6709 para SVM con BoW) y finalmente el modelo de imágenes no logró clasificar las clases correctamente. Como trabajo a futuro, estos últimos modelos de descripción e imagen serán mejorados para construir, junto con los de metainformación, un modelo ensamblado basado en considerables variables del proyecto.

Palabras claves: metainformación, descripción, imagen del proyecto, Máquina de Vectores de Soporte (SVM), Perceptrón Multicapa (MLP), Red Neuronal Convolucional (CNN).

Knowing funding projects destiny has always been the main desire of all entrepreneurs who promote them on the Internet, especially in the technology category because they have the lowest success rates due to their high goals. The present research work was based on building a predictive model whose objective is to estimate the success or failure of technological projects funding in the Kickstarter crowdfunding platform during the duration of its campaign based on its metadata, image and/or description. For this, Machine Learning models (SVM, MLP and CNN) were created for each of these parts. After analyzing all the models with the same metrics mentioned in the tenth antecedent, it was concluded that only metadata models had excellent levels according to their AUC scores (0.8377 for SVM and 0.7043 for MLP), while the description models had a regular performance (0.6746 for SVM with TF-IDF and 0.6709 for SVM with BoW) and finally the image model failed to classify the classes correctly. As future work, these latest models of description and image will be improved to build, together with metadata ones, an assembled model base don considerable project variables.

Palabras claves: project metadata, project description, project image, Support Vector Machine (SVM), Multilayer Perceptron (MLP), Convolutional Neural Network (CNN).

Introducción

Por muchos años, en especial en las dos últimas décadas, diversos proyectos emprendedores han sido lanzados en distintas plataformas web, buscando un objetivo compartido por todos: ser financiados en un determinado plazo para hacer realidad estas ideas. Entre fracasos y éxitos, han surgido nuevas tendencias, así como nuevos enfoques de estudios de estos casos para encontrar la clave que descifre las variables de éxito.

El presente trabajo de investigación se basó formular un modelo ensamblado robusto que determine el estado final de un proyecto, agregando un nuevo enfoque: basarse solamente en proyectos de tecnología, la segunda categoría con más baja probabilidad de éxito al final de una campaña. En estudios previos, los modelos planteados resultaron bastante aceptables debido a que el resto de categorías de proyectos balancearon la inequidad existente en las dos clases del estado.

El reto principal fue el de construir modelos predictivos que consideren las tres características más importantes de un proyecto: la primera basada en la metainformación, la segunda, en la imagen principal del proyecto y la tercera, en la descripción del mismo, para ser ensamblados más adelante en uno solo.

Para ello, se recolectó un total de 27,251 proyectos tecnológicos en Kickstarter entre los periodos 2009-2019, de los cuales 27,035 registros finalmente fueron usados para cada uno de los tres modelos. Algunos proyectos provenientes de países fuera del territorio de los Estados Unidos y en distintos idiomas fueron considerados dentro de esta cantidad ya que no afectó al rendimiento general como en casos particulares de algunos estudios previos.

La principal motivación de este trabajo fue la de aportar una herramienta de ayuda para los emprendedores que les permita estimar el estado final del financiamiento de su proyecto de tecnología, es decir, éxito o fracaso, con un nivel confiable de probabilidad de éxito del mismo durante el transcurso de su campaña, permitiendo además servir de soporte en la toma de decisiones de cara a lograr su principal objetivo.

Capítulo 1

PLANTEAMIENTO DEL PROBLEMA

1.1. Descripción de la Realidad Problemática

El emprendimiento hoy en día es una realidad en todo el mundo. Desde crear productos nuevos hasta crear nuevas formas de hacer las cosas, todo gracias a ideas nacidas a partir de querer satisfacer nuestras propias necesidades.

Nuestro país no es ajeno a ello. El 50.6 % de la población entre 18 y 64 años tiene la expectativa de iniciar un emprendimiento dentro de los tres próximos años de acuerdo al último reporte de Global Entrepreneurship Monitor (GEM) 2014. El 62.3 % de la población entre ese rango de edad, además, tiende a ser más optimista en su percepción de oportunidades. Asimismo, según informa la Cámara de Comercio de Lima, la iniciativa emprendedora responde más a la identificación de una oportunidad de negocio que a una falta de oportunidad de empleo (Redacción Gestión, 2015). Sin embargo, en un estudio más reciente basado en una encuesta realizada a residentes peruanos entre junio y julio del 2017 desarrollada por el equipo GEM Perú y ESAN a 2080 personas entre el mismo rango de edad, el 24.6 % de emprendimientos se encuentra en fase temprana, es decir, representa una dificultad para el emprendedor peruano llegar a etapas más avanzadas como un emprendimiento establecido (negocios con más de 3.5 años, que representan solo el 7.4 % para Perú), ubicando así a nuestro país en la posición 25 de 54 economías a nivel mundial (Redacción Gestión, 2018). En la Figura 1.1 se aprecian algunas ratios del estudio.

Estos resultados desfavorables tienen como base el ecosistema poco beneficioso para los emprendimientos que permitan su establecimiento en el entorno nacional, con condiciones asociadas al acceso de financiamiento, políticas gubernamentales que alienten la implementación de Innovación y Desarrollo en las empresas, acceso a infraestructura física y asesoría

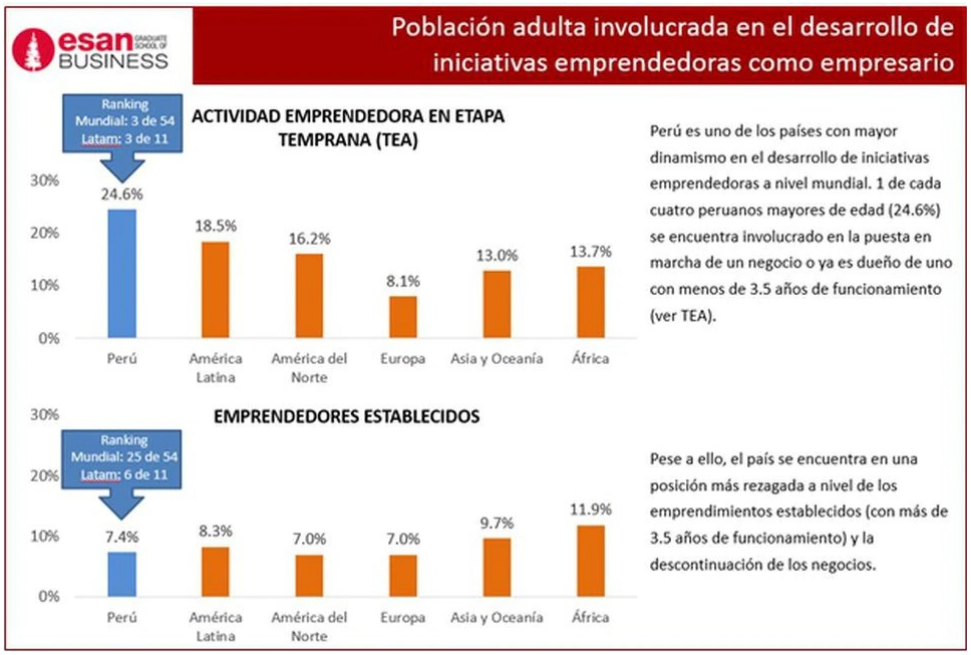

Figura 1.1: Resultados y ratios obtenidos en la encuesta por GEM y ESAN. Fuente: Redacción Gestión, 2018

a nivel comercial y profesional, como sostiene el investigador del equipo GEM Perú Carlos Guerrero (Redacción Gestión, 2018). La Asociación de Emprendedores de Perú (ASEP) afirma, asimismo, que en la región solo se invierte el 1.5 % del PIB en actividades de ciencia, tecnología e innovación, y las limitaciones son dadas por barreras burocráticas ejercidas por el Gobierno y el sector privado (Asociación de Emprendedores de Perú, 2018). En adición a esto, otras razones que representan barreras para emprender son la falta de conocimientos en la iniciación de un negocio, su tramitación, la fuente de financiamiento del proyecto o búsqueda de inversionistas, la cultura, la falta de fomento de emprendimiento y la falta de una red de contactos (Sandoval, s.f.).

Ante estas limitaciones, en la actualidad muchos emprendedores se ven forzados a mostrar sus proyectos al público en la Internet con el fin de captar personas interesadas en ayudarlos en el financiamiento de estos. Por ello, se han creado plataformas web con el fin de permitir la interacción entre los proyectos publicados en un determinado tiempo, el cual puede variar entre 30 y 120 días, y la comunidad en general que desee colaborar con una cantidad de dinero para su financiamiento. El sitio web solo servirá para mostrar los proyectos presentados a detalle por los creadores y la promoción de estos al público. La idea es que, al término de este plazo de tiempo, el proyecto sea financiado y se logre convertir en una realidad. A esta práctica se le conoce como crowdfunding (Universo Crowdfunding, s.f.).

En Latinoamérica, son muy pocos los países los que se incorporan en el crowdfunding,

tales como Chile, México, Argentina y Brasil. Sin embargo, el modelo funciona distinto a países de Norteamérica y Europa debido a la cultura diferente y resistencia a su implementación por la poca confianza en el éxito de los proyectos. En los últimos años se decidió seguir una manera muy similar a los modelos de Estados Unidos, basados en la creación de campañas de un emprendedor para obtener fondos para sus ideas con la moneda norteamericana pero limitados a las leyes económicas de cada país (Solidaridad Latina, s.f.).

Entre los sitios web más conocidos de crowdfunding están Kickstarter e Indiegogo. Kickstarter, desde su inicio en 2009, es una plataforma de financiamiento de proyectos creativos de todo tipo, los cuales incluyen películas, juegos, música, arte, diseño y tecnología. Actualmente, se han registrado más de 162 mil proyectos realizados, 16 millones de contribuyentes y 4,3 miles de millones de dólares fondeados (Kickstarter, s.f.-a). La plataforma utiliza un modelo de financiamiento llamado "todo o nada", el cual consiste en que si un proyecto no alcanza su meta de financiamiento en un determinado plazo de tiempo, no se realiza ninguna transacción de fondos (Kickstarter, s.f.-b). Si bien los patrocinadores apoyan estos proyectos por motivos personales y distintos para hacerlos realidad, ellos no obtienen la propiedad o los ingresos de los proyectos que financian, sino que los creadores conservan la totalidad de su trabajo (Kickstarter, s.f.-c).

Para los proyectos tecnológicos, en contraste, la ratio de éxito es uno de los más bajos de las categorías existentes (20 %) solo por delante de Artesanía y Periodismo, como se aprecia en la Figura 1.2.

Ya existen estudios previos para predecir la probabilidad de éxito de financiamiento para este tipo de proyectos utilizando técnicas de Aprendizaje Automático. Sin embargo, la mayoría de los modelos predictivos propuestos no arrojan resultados con exactitud muy alta ya que su rango varía entre 60 y 70 %. Esto conlleva a generar imprecisión para pronosticar confiablemente el éxito de financiamiento de estos proyectos de tecnología. Para el presente trabajo de tesis, se creó un modelo predictivo alimentado de datos históricos de la plataforma para estimar el estado final de financiamiento de un proyecto aleatorio, así como su probabilidad de éxito.

1.2. Formulación del Problema

Para la formulación de los problemas de la presente investigación, se elaboró un «árbol de problemas» (véase Anexo 1).

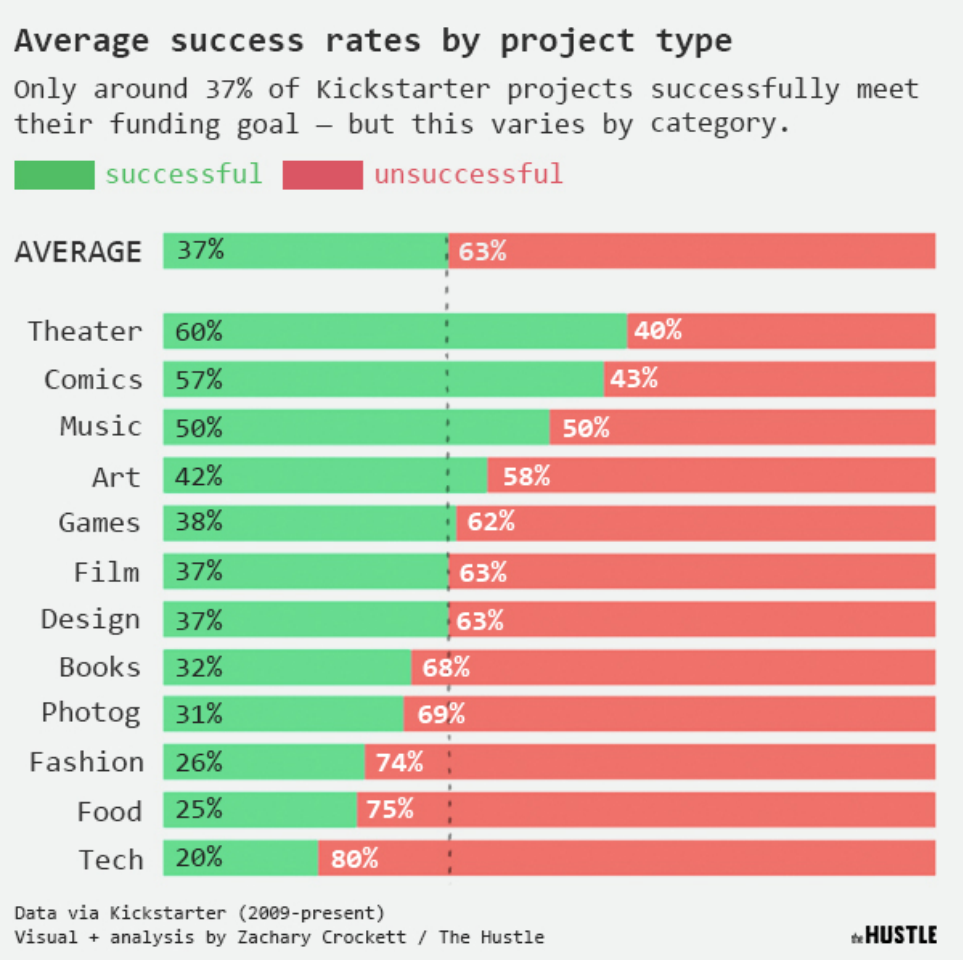

Figura 1.2: Ratio de éxito de proyectos en Kickstarter desde 2009 hasta 2019 (Febrero). Fuente: The Hustle, 2019

1.2.1. Problema General

Bajos niveles de precisión de modelos entrenados de Aprendizaje Automático para cualquier categoría para predecir estado de financiamiento de proyectos de tecnología.

1.2.2. Problemas Específicos

- Variables de proyectos no normalizadas y varianzas altas.
- Datos faltantes o incompletos de proyectos.
- Parámetros de modelos no ajustados.
- Sobreajuste de aprendizaje de modelos y clasificación incorrecta de las dos clases del estado final de financiamiento (exitoso o fracasado).

- Predicción incorrecta de estado de financiamiento de un proyecto tecnológico.

1.3. Objetivos de la Investigación

Para la formulación de los objetivos de la presente investigación, se elaboró un «árbol de objetivos» (véase Anexo 2)

1.3.1. Objetivo General

Construir modelo(s) de Aprendizaje Automático entrenado(s) para predecir correctamente proyectos de tecnología con nivel de precisión aceptable.

1.3.2. Objetivos Específicos

- Normalizar variables de proyectos y reducir niveles altos de varianza.
- Eliminar datos faltantes o incompletos de proyectos.
- Ajustar parámetros de modelos.
- Evitar sobreajuste de aprendizaje de modelos.
- Predecir correctamente el estado final de financiamiento de cualquier proyecto tecnológico (éxito o fracaso).

1.4. Justificación de la Investigación

1.4.1. Teórica

Esta investigación se basa en crear un modelo de Aprendizaje Automático que sea aplicable a proyectos de tecnología de la plataforma Kickstarter por presentar bajas performances en antecedentes.

1.4.2. Práctica

Al culminar la investigación, se ofrecerá un modelo predictivo confiable que ayude a los emprendedores en la toma de decisiones respecto a sus proyectos a partir del insight obtenido de los resultados que deriven a la manipulación de los datos de entrada.

1.4.3. Metodológica

Se creará un modelo predictivo a partir de las variables finales seleccionadas, previa limpieza de datos. Luego, será entrenado y evaluado por las métricas correspondientes. Finalmente, se lanzará una versión de prueba que reciba datos de entrada para predecir la viabilidad de un proyecto de tecnología.

1.5. Delimitación del Estudio

1.5.1. Espacial

Para la presente investigación, se considerará el territorio de los Estados Unidos ya que tanto la campaña del proyecto a servir para la investigación como los datos fuentes de proyectos relacionados financiados previamente, que servirán para la elaboración del modelo predictivo, se encuentran en dicho país.

1.5.2. Temporal

El periodo de tiempo abarcará desde el año 2009, fecha en el cual se tiene registrado los primeros conjuntos de datos de proyectos en Kickstarter hasta el mes de agosto del año 2019, últimos registros descargados hasta el inicio del presente trabajo.

1.5.3. Conceptual

La presente investigación consistirá en la implementación de un modelo predictivo del estado de financiamiento de un proyecto tecnológico en Kickstarter basado en técnicas y conceptos de Aprendizaje Automático, previamente evaluando cuál de todas las existentes genera un mejor desempeño para su uso y análisis de resultados.

1.6. Hipótesis

1.6.1. Hipótesis General

El modelo entrenado de Aprendizaje Automático logrará predecir correctamente proyectos de tecnología con nivel de precisión aceptable.

1.6.2. Hipótesis Específicas

- Las variables de los proyectos descargados se normalizarán y se reducirán los niveles altos de varianza.
- Los datos faltantes o incompletos de los proyectos serán eliminados.
- Los parámetros de los modelos usados serán ajustados.
- Se evitará el sobreajuste de aprendizaje de modelos para clasificar correctamente las dos clases del estado final de financiamiento.
- El estado final de financiamiento de cualquier proyecto tecnológico será predicho correctamente.

1.6.3. Matriz de Consistencia

A continuación se presenta la matriz de consistencia elaborada para la presente investigación (véase Anexo A.1).

Capítulo 2

MARCO TEÓRICO

2.1. Antecedentes de la investigación

En esta sección se presentarán diversos trabajos de investigación basados en la predicción de éxito o fracaso de campañas en Kickstarter o plataformas similares y el análisis de estas utilizando conjunto de datos de la propia plataforma o almacenadas en otros repositorios, con sus variables respectivas. En la mayoría de casos consideraron variables básicas que se obtienen del repositorio de las plataformas de crowdfunding, en otros casos consideraron variables cualitativas basadas en texto y descripción de proyectos, y en otros antecedentes usaron nuevas técnicas poco convencionales como el Aprendizaje Profundo e híbridos de modelos para obtener los mejores resultados posibles. Asimismo, a continuación se presenta un cuadro resumen (véase Anexo B.1) de lo que se presenta en esta sección.

2.1.1. Primer antecedente: «Supervised Learning Model For Kickstarter Campaigns With R Mining» (Kamath & Kamat, 2018)

Kamath y Kamat realizaron un artículo de investigación el cual fue publicado en la revista «Resources Policy» en el año 2018. Este fue titulado «Supervised Learning Model For Kickstarter Campaigns With R Mining» la cual traducida al español significa «Estimación del precio del cobre utilizando el algoritmo bat».

2.1.1.1. Planteamiento del Problema y objetivo

hhhhj

2.1.1.2. Técnicas empleadas por los autores

Los autores plantearon emplear una combinación entre la función de series de tiempo y el aljhkk.

2.1.1.3. Metodología empleada por los autores

gfhhhh

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}\left(O_i - T_i\right)^2}{N}} \qquad \text{(Ecuación 2.1)}$$

gfghf tal forma mejorar aún más la precisión de la predicción del precio del cobre.

2.1.1.4. Resultados obtenidos

Las funciones de serie de tiempo más importantes se usaron para estimar los cambios en el precio del cobre. Entre ellos, la serie BMMR con una media de RMSE de 0.449 presentó la mejor estimación. El algoritmo Bat se usó para modificar la función de tiempo BMMR debido a su alta capacidad para estimar los cambios en el precio del metal. Se obtuvo un RMSE de 0.132 de la ecuación modificada con BA. Los resultados obtenidos tienen una precisión mucho mayor y, a diferencia del BMMR, están más cerca de la realidad.

2.1.2. Segundo antecedente: «Predicting Success in Equity Crowdfunding» (Beckwith, 2016)

Beckwith realizó un artículo de investigación el cual fue publicado en la revista «Resources Policy» en el año 2018. Este fue titulado «Predicting Success in Equity Crowdfunding» la cual traducida al español significa «Estimación del precio del cobre utilizando el algoritmo bat».

2.1.2.1. Planteamiento del Problema y objetivo

hhhhj

2.1.2.2. Técnicas empleadas por los autores

2.1.2.3. Metodología empleada por los autores

2.1.2.4. Resultados obtenidos

2.1.3. Tercer antecedente: «Money Talks: A Predictive Model on Crowdfunding Success Using Project Description» (Zhou, Zhang, Wang, Du, Qiao & Fan, 2018)

Zhou y col. realizaron un artículo de investigación el cual fue publicado en la revista «Resources Policy» en el año 2018. Este fue titulado «Money Talks: A Predictive Model on Crowdfunding Success Using Project Description» la cual traducida al español significa «Estimación del precio del cobre utilizando el algoritmo bat».

2.1.3.1. Planteamiento del Problema y objetivo

2.1.3.2. Técnicas empleadas por los autores

2.1.3.3. Metodología empleada por los autores

2.1.3.4. Resultados obtenidos

2.1.4. Cuarto antecedente: «The Determinants of Crowdfunding Success: A Semantic Text Analytics Approach» (Yuan, Lau & Xu, 2016)

Yuan y col. realizaron un artículo de investigación el cual fue publicado en la revista «Resources Policy» en el año 2018. Este fue titulado «The Determinants of Crowdfunding Success: A Semantic Text Analytics Approach» la cual traducida al español significa «Estimación del precio del cobre utilizando el algoritmo bat».

2.1.4.1. Planteamiento del Problema y objetivo

2.1.4.2. Técnicas empleadas por los autores

2.1.4.3. Metodología empleada por los autores

2.1.4.4. Resultados obtenidos

2.1.5. Quinto antecedente: «Will your Project get the Green light? Predicting the success of crowdfunding campaigns» (Chen, Chen, Chen, Yang & Lin, 2015)

Chen y col. realizaron un artículo de investigación el cual fue publicado en la revista «Resources Policy» en el año 2018. Este fue titulado «Will Your Project Get the Green Light? Predicting the Success of Crowdfunding Campaigns» la cual traducida al español significa «Estimación del precio del cobre utilizando el algoritmo bat».

2.1.5.1. Planteamiento del Problema y objetivo

2.1.5.2. Técnicas empleadas por los autores

2.1.5.3. Metodología empleada por los autores

2.1.5.4. Resultados obtenidos

2.1.6. Sexto antecedente: «Project Success Prediction in Crowdfunding Environments» (Li, Rakesh & Reddy, 2016)

Li y col. realizaron un artículo de investigación el cual fue publicado en la revista «Resources Policy» en el año 2018. Este fue titulado «Project Success Prediction in Crowdfunding Environments» la cual traducida al español significa «Estimación del precio del cobre utilizando el algoritmo bat».

2.1.6.1. Planteamiento del Problema y objetivo

2.1.6.2. Técnicas empleadas por los autores

2.1.6.3. Metodología empleada por los autores

2.1.6.4. Resultados obtenidos

2.1.7. Séptimo antecedente: «Effect of Social Media Connectivity on Success of Crowdfunding Campaigns» (Kaur & Gera, 2017)

Kaur y Gera realizaron un artículo de investigación el cual fue publicado en la revista «Resources Policy» en el año 2018. Este fue titulado «Effect of Social Media Connectivity on Success of Crowdfunding Campaigns» la cual traducida al español significa «Estimación del precio del cobre utilizando el algoritmo bat».

2.1.7.1. Planteamiento del Problema y objetivo

2.1.7.2. Técnicas empleadas por los autores

2.1.7.3. Metodología empleada por los autores

2.1.7.4. Resultados obtenidos

2.1.8. Octavo antecedente: «Prediction of Crowdfunding Project Success with Deep Learning» (Yu, Huang, Yang, Liu, Li & Tsai, 2018)

Yu y col. realizaron un artículo de investigación el cual fue publicado en la revista «Resources Policy» en el año 2018. Este fue titulado «Prediction of Crowdfunding Project Success with Deep Learning» la cual traducida al español significa «Estimación del precio del cobre utilizando el algoritmo bat».

2.1.8.1. Planteamiento del Problema y objetivo

2.1.8.2. Técnicas empleadas por los autores

2.1.8.3. Metodología empleada por los autores

2.1.8.4. Resultados obtenidos

2.1.9. Noveno antecedente: «Estimating the Days to Success of Campaigns in Crowdfunding: A Deep Survival Perspective» (Jin, Zhao, Chen, Liu & Ge, 2019)

Jin y col. realizaron un artículo de investigación el cual fue publicado en la revista «Resources Policy» en el año 2018. Este fue titulado «Estimating the Days to Success of Campaigns in Crowdfunding: A Deep Survival Perspective» la cual traducida al español significa «Estimación del precio del cobre utilizando el algoritmo bat».

2.1.9.1. Planteamiento del Problema y objetivo

2.1.9.2. Técnicas empleadas por los autores

2.1.9.3. Metodología empleada por los autores

2.1.9.4. Resultados obtenidos

2.1.10. Décimo antecedente: «Success Prediction on Crowdfunding with Multimodal Deep Learning» (Cheng, Tan, Hou & Wei, 2019)

Cheng y col. realizaron un artículo de investigación el cual fue publicado en la revista «Resources Policy» en el año 2018. Este fue titulado «Success Prediction on Crowdfunding with Multimodal Deep Learning» la cual traducida al español significa «Estimación del precio del cobre utilizando el algoritmo bat».

2.1.10.1. Planteamiento del Problema y objetivo

2.1.10.2. Técnicas empleadas por los autores

2.1.10.3. Metodología empleada por los autores

2.1.10.4. Resultados obtenidos

2.2. Bases Teóricas

2.2.1. Inteligencia Artificial

La Inteligencia Artificial es la inteligencia llevado a cabo por máquinas en las que una máquina "inteligente" ideal es un agente flexible que percibe su entorno y lleva a cabo acciones que maximicen sus posibilidades de éxito en algún objetivo (Poole y col., 1998). Este término se aplica cuando una máquina imita las funciones "cognitivas" que asocian los humanos con otras mentes (Russell & Norvig, 2009).

Durante la historia de la humanidad, se han seguido 4 enfoques: dos centrados en el comportamiento humano y dos enfocados en torno a la racionalidad. El enfoque centrado en el comportamiento humano se basa en una ciencia empírica, es decir, mediante experimentos que incluyen hipótesis y confirmaciones. Este enfoque nace a partir de la prueba de Alan Turing, en 1950, en la cual, el célebre matemático inglés diseñó una prueba basada en la incapacidad de diferenciar entre entidades inteligentes indiscutibles y seres humanos por parte de un computador. Si este era capaz de diferenciar y superar la prueba mientras que el humano no, se afirma que se trataba de una "máquina inteligente". Por ello, el computador debía contar con las siguientes capacidades: procesamiento de lenguaje natural para poder comunicarse, representación del conocimiento describiendo lo que percibe de su entorno, razonamiento automático utilizando la información procesada en su interior, y aprendizaje automático para adaptarse a nuevos eventos. Si el evaluador decide incluir una señal de video para evaluar la percepción de la computadora, se dice que se está realizando la Prueba Global de Turing. Para superarla, además de las 4 anteriormente mencionadas, la computadora debe contar además con las capacidades de visión computacional para percibir objetos y robótica con el fin de manipularlos. Todas estas seis capacidades o disciplinas abarcan la mayor parte de la Inteligencia Artificial (Russell & Norvig, 2004).

Por el otro lado, el enfoque racional implica una combinación de ingeniería y matemáticas basándose en las "leyes del pensamiento". Estas parten de la Grecia antigua, planteadas por

grandes filósofos como Aristóteles en su intento de codificar la "manera correcta de pensar", lo que más adelante derivó al estudio de la lógica. Más adelante, en el siglo XIX, se construyeron programas capaces de resolver problemas en notación lógica. De ahí que la tradición logista dentro del campo de la Inteligencia Artificial trata de construir sistemas inteligentes con estas capacidades. De todo lo anterior dicho respecto al enfoque racional se creó el término de un agente racional, el cual actúa intentando lograr el mejor resultado, o de existir incertidumbre, el mejor resultado esperado. Finalmente, la amplia aplicación de la Inteligencia Artificial y sus fundamentos derivan en muchas ciencias de las cuales se pueden mencionar, además de la filosofía y las matemáticas, a la economía, neurociencia, psicología, la ingeniería computacional, la teoría de control y cibernética, y hasta la lingüística (Russell & Norvig, 2004).

Pero, ¿cómo es surge este amplio estudio de la Inteligencia Artificial? En 1943, basándose en la fisiología básica y funcionamiento de las neuronas en el cerebro, el análisis formal de la lógica proposicional de Russell y Whitehead, y la teoría computacional de Turing, dos estudiosos en neurociencia realizaron juntos el que sería considerado primer trabajo de Inteligencia Artificial. Warren McCulloch y Walter Pitts propusieron un modelo constituido por neuronas artificiales, en el que cada una de ellas se caracterizaba por estar "activada" o "desactivada"; la del primer tipo daba como resultado a la estimulación producida por una cantidad suficiente de neuronas vecinas. Como ejemplo, mostraron que cualquier función de cómputo podría calcularse mediante alguna red de neuronas interconectadas y que todos los conectores lógicos eran capaces de ser implementados usando estructuras sencillas de red. Seis años más adelante, Donald Hebb propuso una regla de actualización de intensidades de conexiones entre las neuronas, la que actualmente se le conoce como la "regla de aprendizaje Hebbiano" vigente hasta nuestros días. En 1956, Allen Newell y Herbert Simon inventaron un programa de computación en el taller de Dartmouth de John McCarthy, que era capaz de pensar de forma no numérica, basado en el Teórico Lógico, artículo que, además, fue rechazado de ser publicado en la revista *Journal of Symbolic Logic*. A pesar de ello, los trabajos de los colaboradores presentes en dicho taller se mantuvieron por 20 años más, siendo McCarthy quien acuñó el término de "Inteligencia Artificial" a este campo (Russell & Norvig, 2004).

En la década de los años 80, la Inteligencia Artificial dio el gran salto de formar parte de la industria, en especial, de las compañías más grandes de los países desarrollados a través de grupos especializados para la realización de investigaciones de sistemas expertos, así como en la construcción de computadoras cada vez más potentes y capaces de resolver tareas más complejas.

Actualmente, la IA cuenta con muchas aplicaciones como la Minería de Datos, el procesamiento de lenguaje natural, la robótica, los videojuegos, entre otros. Dentro de ella se pueden encontrar otras ramas como por ejemplo el Aprendizaje Automático, Visión computacional,

etcétera.

2.2.2. Aprendizaje Automático

El Aprendizaje Automático (*Machine Learning* por su nombre en inglés) es una rama de la Inteligencia Artificial cuyo fin es desarrollar técnicas que las computadoras pueden aprender a través de encontrar algoritmos y heurísticas que conviertan muestras de datos en programas sin necesidad de hacerlos (Russell & Norvig, 2009). Sus algoritmos están compuestos por muchas tecnologías, como por ejemplo Aprendizaje Profundo, Redes Neuronales y Procesamiento de lenguaje natural, utilizadas en el aprendizaje supervisado y no supervisado, las cuales operan guiadas por lecciones de información existente (Gartner, 2019). La premisa básica del aprendizaje automático es construir algoritmos que puedan recibir datos de entrada y usar análisis estadísticos para predecir una salida mientras se actualizan las salidas a medida que se dispone de nuevos datos (Alpaydin, 2014).

Los tres tipos de aprendizaje principales son:

- **Aprendizaje supervisado**: Se trabajan con datos etiquetados buscando obtener una función que asigne una respuesta de salida adecuada, denominadas etiquetas, a partir de unos datos de entrada denominadas características (Zambrano, 2018). Por lo general, los datos de entrada son conocidos como variables dependientes o X, mientras que los datos de salida son llamadas variables independientes o Y. Se le dice supervisado ya que el resultado depende de los datos que recibe de entrada, afectando su performance si estos son alterados.

  Existen dos tipos de aprendizaje supervisado. El primero es la regresión, que consiste en obtener como resultado un número específico a partir de un conjunto de variables de las características, representado en la Figura 2.1; mientras que por otra parte está la clasificación, el cual se basa en encontrar distintos patrones ocultos para clasificar los elementos del conjunto de datos en diferentes grupos, como se aprecia en la Figura 2.2 (Zambrano, 2018).

  Para el segundo tipo de aprendizaje supervisado, el algoritmo más usado es el de los K Vecinos más cercanos o k-NN Nearest Neighbour en inglés. Este se basa en la idea de que los nuevos ejemplos serán clasificados a la clase a la cual pertenezca la mayor cantidad de vecinos más cercanos del conjunto de entrenamiento más cercano a él. Sin embargo, el número k de vecinos más cercanos lo decide el usuario, de preferencia impar, para evitar ambigüedad al momento de clasificar un registro por parte del algoritmo (esto puede ocasionarse por las mismas distancias existentes entre dos o más registros). Otra variante

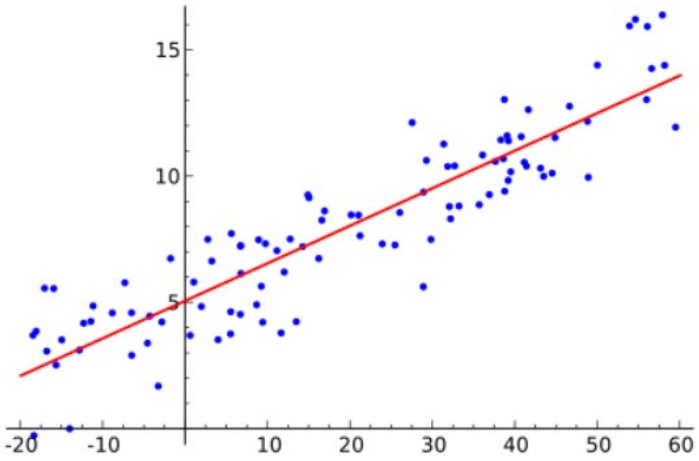

Figura 2.1: Ejemplo de algoritmo de regresión. Fuente: Zambrano, 2018

aplicada consiste en la ponderación de cada vecino de acuerdo a la distancia entre él y el ejemplar a ser clasificado, asignando mayor peso a los más próximos (Sancho Caparrini, 2018). Por ejemplo, si x es el ejemplo que se desea clasificar, V son las posibles clases de clasificación, yxi es el conjunto de los k ejemplos de entrenamiento más cercano, se define la siguiente fórmula:

y finalmente, la clase asignada a x es aquella que verifique que la suma de los pesos de sus representantes sea la máxima, representándose en la Figura 2.3:

- **Aprendizaje no supervisado**: A diferencia de la anterior, aquí se trabaja con datos no etiquetados para entrenar el modelo, ya que el fin es de carácter exploratorio y descriptivo de la estructura de los datos. No existen variables independientes o Y.

  La función es agrupar ejemplares, por lo que el algoritmo los cataloga por similitud en sus características y a partir de ahí, crea grupos o clústeres sin tener la capacidad de definir cómo es cada individualidad de cada uno de los integrantes de los mismos (Zambrano, 2018).

  El algoritmo usado para este tipo de aprendizaje es el de las K medias o k-means en inglés. Este intenta encontrar una partición de las muestras en K agrupaciones, de manera que cada ejemplar pertenezca a una de ellas de acuerdo al centroide más cercano. Si bien el valor de K es definido por el usuario, a partir de pruebas de varias iteraciones se le puede consultar al algoritmo cuál es su valor óptimo. La intención es minimizar la varianza total del sistema. Por ejemplo, si se tiene el centroide ci de la agrupación i-ésima, yxji es el conjunto de ejemplos clasificados en esa agrupación, la función para lograr esto es la siguiente:

  Representándose en la Figura 7, los pasos seguidos para este algoritmo comienzan con

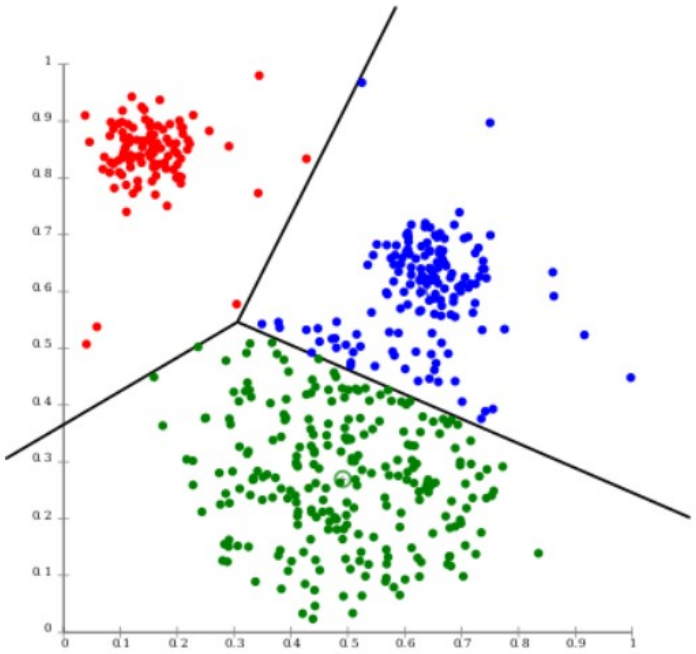

Figura 2.2: Ejemplo de algoritmo de clasificación. Fuente: Zambrano, 2018

la selección de los K puntos como como centros de los grupos. Luego, se asignarán los ejemplos al centro más cercano y se calculará el centroide de los ejemplos asociados a cada grupo. Finalmente, estos dos últimos pasos se repetirán hasta que ninguno de los centros pueda ser reasignados en las iteraciones.

- **Aprendizaje por refuerzo**: Se basa en que un agente racional puede tomar una decisión a partir de una retroalimentación llamada recompensa o refuerzo. A diferencia del Aprendizaje Supervisado, en donde el agente puede aprender solamente a partir de ejemplos dados, en este caso no basta solamente con proporcionárselos sino también de "informarle" si lo está haciendo de la manera correcta o no. Por ejemplo, un agente que intenta aprender a jugar ajedrez necesita saber que algo bueno ha ocurrido cuando gana y algo malo ha ocurrido cuando pierde. La mejor recompensa que busca al finalizar el juego es vencer al oponente, y para ello debe estudiar todos los movimientos que este haga, la posición de las fichas en el tablero, entre otros. A este conjunto se le conoce como entorno o medio ambiente (Russell & Norvig, 2004). Entonces, en resumen, representando en la Figura 8, y mencionando otro ejemplo, el aprendizaje por refuerzo está compuesto por un agente (Pacman) en un estado determinado (su ubicación o posición actual) dentro de un medio ambiente (el laberinto). La recompensa positiva que busca Pacman son los puntos por comer, mientras que la negativa será la de morir si se cruza con un fantasma, en base a la acción (desplazamiento a un nuevo estado) que realice .

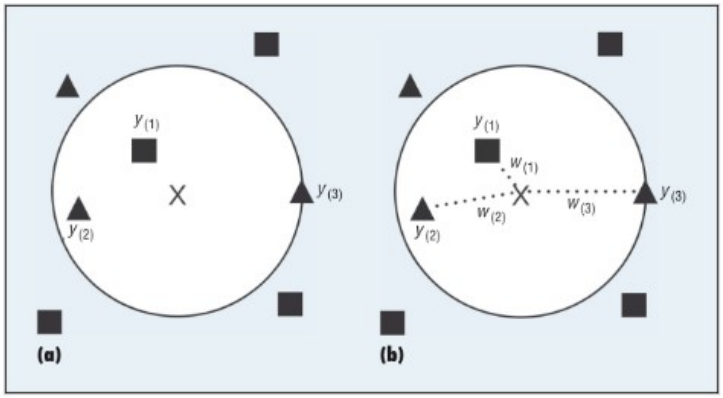

Figura 2.3: Algoritmo de K Vecinos más cercanos con pesos ponderados. Fuente: Sancho Caparrini, 2018

2.2.3. Aprendizaje Profundo

El Aprendizaje Profundo (*Deep Learning* por su nombre en inglés) es un tipo de Aprendizaje Automático que entrena a una computadora para que realice tareas como las realizadas por los seres humanos, desde la identificación de imágenes hasta realizar predicciones y reconocer el lenguaje humano. El Aprendizaje Profundo configura parámetros básicos acerca de los datos y entrena a la computadora para que aprenda por su cuenta reconociendo patrones mediante el uso de múltiples capas de procesamiento . Se basa en teorías acerca de cómo funciona el cerebro humano .

La principal diferencia con el Aprendizaje Automático es que el Aprendizaje Profundo se basa en la extracción de características y clasificación al mismo tiempo luego de recibir una entrada, algo que en la primera técnica ocurre por separado, como se aprecia en la Figura 9.

Por un lado, mientras en el aprendizaje automático o de máquina, el ordenador extrae conocimiento a través de experiencia supervisada, en el aprendizaje profundo está menos sometido a supervisión. Mientras que el primer tipo de aprendizaje consume muchísimo tiempo y se basa en proponer abstracciones que permiten aprender al ordenador, en el segundo no consume demasiado tiempo y por el contrario de su par, crea redes neuronales a gran escala que permiten que el ordenador aprenda y piense por sí mismo sin necesidad directa de intervención humana. Actualmente, el aprendizaje profundo se usa para crear softwares capaces de determinar emociones o eventos descritos en textos, reconocimiento de objetos en fotografías y realizar predicciones acerca del posible comportamiento futuro de las personas. Empresas como Google (proyecto Google Brain) o Facebook (Unidad de investigación en IA) han puesto en marcha proyectos basados en esta rama para potenciar y mejorar sus algoritmos con el fin de ofrecer una mejor experiencia de sus servicios a sus clientes .

2.2.4. Modelo Predictivo

Son modelos de datos estadísticos utilizados para predecir el comportamiento futuro. En estos, se recopilan datos históricos y actuales, se formula un modelo estadístico, se realizan predicciones y el modelo se valida a medida que se dispone de datos adicionales. Los modelos predictivos analizan el rendimiento pasado para evaluar la probabilidad de que un cliente muestre un comportamiento específico en el futuro. En esta categoría también abarca la búsqueda de patrones ocultos .

2.2.5. Minería de Datos

La Minería de Datos es un campo de la estadística y las ciencias de la computación referido al proceso que intenta descubrir patrones en grandes volúmenes de conjuntos de datos . Normalmente, estos patrones no pueden detectarse mediante la exploración tradicional de datos porque sus relaciones son demasiado complejas o por su gran volumen. Para ello, utiliza métodos de Inteligencia Artificial, Aprendizaje Automático, estadística y sistemas de bases de datos. Estos patrones son recopilados y definidos como un modelo de minería de datos, los cuales pueden aplicarse en los siguientes escenarios :

- Previsión.
- Riesgo y probabilidad.
- Recomendaciones.
- Buscar secuencias.
- Agrupación.

La generación de un modelo de minería de datos forma de un macro-proceso descrita en los siguientes seis pasos representados en la Figura 10:

2.2.6. Metodologías de Minería de Datos

Dentro de los sistemas de analítica de negocio, Big Data y Minería de Datos, las tres metodologías más usadas se encuentran CRISP-DM, SEMMA y KDD .

- **CRISP-DM** (Cross Industry Standard Process for Data Mining):

  Esta metodología presenta seis fases representadas en la Figura 11 a continuación.

- En la comprensión del negocio se determinan los objetivos y requerimientos desde el lado del negocio, así como generar plan del proyecto.
- En la comprensión de los datos se logra entender el significado de las variables existentes, así como el entendimiento de los datos desde su recopilación hasta su verificación de calidad.
- En la preparación de los datos se prepara el conjunto de datos adecuado que servirán para la construcción del modelo. Por ello, la calidad de los datos es un factor relevante y ello requiere la exclusión de redundancia y valores que no ayuden a establecer buena comprensión y resultados más adelante. A esto se le conoce como limpieza de datos.
- En el modelado se aplican técnicas de minería de datos en el conjunto de datos creado en el paso anterior. Para ello, se evalúan entre varias la que mejor performance desempeñe y luego se construye el o los modelos que busquen determinar un objetivo.
- En la evaluación se evalúan los posibles modelos del paso anterior a partir del nivel de importancia de acuerdo a las necesidades del negocio y performance que estos cuentan.
- El despliegue, finalmente, utiliza el modelo final creado para determinar los objetivos que se buscan cumplir en los requerimientos y ayudar en la toma de decisiones.

- **SEMMA** (Sample – Explore – Modify – Model – Assess):

  Esta metodología cuenta con cinco fases como se aprecia en la Figura 12. A diferencia de la anterior, esta metodología se enfoca más en el modelado.

  - En la Muestra (Sample) se crea una muestra significativa.
  - En la Exploración (Explore) se comprenden los datos con el fin de encontrar relaciones entre variables y anomalías.
  - En la Modificación (Modify) se transforman las variables para las necesidades del modelo.
  - En la Modelización (Model) se aplican uno o varios modelos sobre el conjunto de datos para buscar resultados.
  - En el Asesoramiento (Assessment) se evalúan los resultados obtenidos del modelo.

- **KDD** (Knowledge Discovery and Data Mining):

Esta metodología se refiere al proceso de encontrar conocimiento alguno en el dato y, a diferencia de sus predecesores, se enfoca en crear aplicaciones de minería de datos. Consta de cinco fases más 1 previa y 1 posterior basadas en la generación de conocimiento como se muestra en la Figura 13.

- En la fase Pre KDD se comprende el dominio del negocio, así como también se identifican las necesidades del cliente.
- En la selección, primero se identifica el conjunto de datos a usar y luego se seleccionan la muestra y las variables para la exploración.
- En el pre-procesamiento, se realiza la limpieza de datos y se elimina el ruido, así como los valores atípicos.
- En la transformación se implementan métodos de reducción de dimensiones para reducir el número de variables efectivas.
- En la Minería de datos, se elige el tipo de tarea de minería de datos (clasificación, regresión, agrupamiento, entre otros) así como el algoritmo, los métodos, los modelos y parámetros apropiados.
- En la interpretación y evaluación se analizan los resultados dados.
- En la fase Post KDD finalmente se consolida el conocimiento adquirido.

Luego de presentar las tres metodologías más usadas, la pregunta dada es ¿cuál de los tres representa la mejor opción para usar? Las tres metodologías tienen distinto número de pasos, así como distintos enfoques, tal cual se observa en el siguiente resumen de la Tabla 2.

Sin embargo, la elección depende de los involucrados que finalmente usarán el modelo en el negocio. La mayoría de investigadores siguen la metodología KDD debido a que es más completo y su exactitud. Para aquellos objetivos enfocados más en la compañía como la integración usada por SAS Enterprise Miner con su software se utilizan SEMMA y CRISP-DM. Esta última resulta ser más completa de acuerdo a los estudios.

2.2.7. Técnicas de Minería de Datos

Existe una gran variedad de técnicas para la Minería de Datos. Las más importantes y utilizadas en los antecedentes de la investigación se mencionan a continuación (Microsoft, 2018).

- **Redes Neuronales Artificiales** (RNA): Es un sistema de computación que consiste en un número de elementos o nodos simples, pero altamente interconectados, llamados “neuro-

nas", que se organizan en capas que procesan información utilizando respuestas de estado dinámico a entradas externas (poner ref).

Este sistema de programas y estructura de datos se aproxima al funcionamiento del cerebro humano. Una red neuronal implica tener un gran número de procesadores funcionando en paralelo, teniendo cada uno de ellos su propia esfera de conocimiento y acceso a datos en su memoria local. Normalmente, una se alimenta con grandes cantidades de datos y un conjunto dado de reglas acerca de las relaciones. Luego, un programa puede indicar a la red cómo debe comportarse en respuesta a un estímulo externo o si puede iniciar la actividad por sí misma (poner ref).

Para entender mejor cómo funciona una red neuronal, hay que describir qué es una neurona. Una neurona es una célula del cerebro cuya función principal es la recogida, procesamiento y emisión de señales eléctricas. Debido a que se piensa que la capacidad de procesamiento de información del cerebro proviene de redes de este tipo de neuronas, los primeros trabajos en Inteligencia Artificial se basaron en crear redes neuronales artificiales para emular este comportamiento, en 1943 con un modelo matemático, mostrado en la Figura 14, por los ya mencionados anteriormente McCulloch y Pitts. Estos y posteriores trabajos potenciaron lo que hoy en día se conoce como el campo de la neurociencia computacional (poner ref). Años más tarde, en 1958, se desarrolló el concepto del perceptrón por Rosenblatt, el cual tenía la capacidad de aprender y reconocer patrones sencillos, formado por entradas, neurona, función de adaptación (sigmoidal, tangencial, en escalón, etc.) y salida.

La última figura descrita muestra, además de los pesos, funciones de activación tanto para la entrada (aj) como para la salida (ai). Pero, ¿qué son estas funciones y para qué sirven?

Para comenzar, las redes neuronales están compuestas de nodos (la elipse) conectados a través de conexiones dirigidas (las flechas). Una conexión del nodo j a la unidad i sirve para propagar la activación aj de j a i. Asimismo, cada conexión tiene un peso numérico W(j,i) que determina la fuerza y el signo de la conexión. Para calcular cada nodo i, se realiza una suma ponderada de sus entradas (producto entre pesos y nodos de entrada j), y se le añade el sesgo (bias) teta-i (aumenta/disminuye el valor de la combinación lineal de las entradas):

Posteriormente, se efectúa una función de activación g a esta suma para producir la salida:

Entonces, aquí se explica los dos objetivos de una función de activación. En primer lugar, se desea que el nodo esté "activo" (cercano a +1) cuando las entradas correctas sean dadas, e "inactiva" (cercano a 0) cuando las entradas erróneas sean proporcionadas. En segundo lugar, la activación tiene que ser no lineal porque, de lo contrario, la red neuronal

colapsaría en su totalidad con una función lineal sencilla, como se aprecia en el ejemplo de la Figura 15.

Entre las funciones de activación que más destacan son las siguientes:

- **Función sigmoide o logística**: Toma los valores de entrada que oscilan entre infinito negativo y positivo, y restringe los valores de salida al rango entre 0 y 1. Frecuentemente es usada en Redes Multicapa (MLP) entrenadas con el algoritmo de propagación inversa. Se representa como en la Figura 16 y su fórmula para calcular su nuevo valor es:

  Un dato curioso de esta función relacionado con la regresión logística es que el nombre de esta última no deriva de una regresión. Por el contrario, se debe a que, al principio de la neurona, se realiza una combinación lineal muy parecida a una regresión lineal y después se aplica la función logística o sigmoide. De ahí el origen del nombre (poner ref).

  - **Regresión Logística**: Como se mencionó antes, es similar a un modelo de regresión lineal, pero está adaptado para modelos en los que la variable dependiente es dicotómica, es decir, presenta solo dos posibles valores. Resulta muy útil para los casos en los que se desea predecir la presencia o ausencia de una característica o resultado según los valores de un conjunto de predictores (poner ref). Su función de coste que se optimiza con gradiente descendiente se representa mediante la siguiente fórmula:
    Donde la primera parte de la ecuación está conformada por el logaritmo de la probabilidad de éxito y la segunda, por la de fracaso.
  - **Gradiente descendiente**: Es un método de optimización numérica para estimar los mejores coeficientes, fundamental en Deep Learning para entrenar redes neuronales y en muchos casos, para la regresión logística, siendo mejor que el método de mínimos cuadrados (poner ref). A través de una función E(W), proporciona el error que comete la red en función del conjunto de pesos sinápticos W. El objetivo del aprendizaje será encontrar la configuración de pesos que corresponda al mínimo global de la función de error o coste (poner ref).
    En general, la función de error es una función no lineal, por lo que el algoritmo realiza una búsqueda a través del espacio de parámetros que, se aproxime de forma iterada a un error mínimo de la red para los parámetros adecuados, como se aprecia en la Figura 17 (poner ref).
    El Descenso del Gradiente, como también se le conoce, es el algoritmo de entrenamiento más simple y también el más extendido y conocido. Solo hace

uso del vector gradiente, y por ello se dice que es un método de primer orden (poner ref). Un gradiente es la generalización de la derivada. Matemática, la derivada de una función mide la rapidez con la que cambia el valor de esta, según varié el valor de su variable independiente. La gradiente se calcula con derivadas parciales, por lo que al actualizar los coeficientes W para un tiempo t, se usa la regla (poner ref).

Donde alfa es el "ratio de aprendizaje", el cual controla el tamaño de la actualización, si este es demasiado grande será más difícil encontrar los coeficientes que minimicen la función de coste o error; la actualización de W es proporcional al gradiente; y se usa la resta para ir en dirección opuesta al gradiente como en la Figura 18.

- **Propagación hacia atrás**: También conocido en inglés como Backpropagation, es un método que consta de dos fases: en la primera se aplica un patrón, el cual se propaga por las distintas capas que componen la red hasta producir la salida de la misma. Luego, esta se compara con la salida deseada y se calcula el error cometido por cada neurona de salida. Estos errores se transmiten hacia atrás, partiendo de la capa de salida, hacia todas las neuronas de las capas intermedias [Fritsch, 1996] (poner ref). La actualización iterativa de los pesos que el algoritmo propone es mediante la siguiente fórmula: Para entender mejor la teoría y la fórmula de actualización de pesos, se seguirá el siguiente ejemplo del conjunto de redes de la Figura 20.

  Se tiene una red neuronal con tres nodos de entradas (x1=1, x2=4 y x3=5) con dos pesos respectivos cada una (W1=0.1 y W2=0.2 para x1; W3=0.3 y W4=0.4 para x2; W5=0.5 y W6=0.6 para x3), dos capas ocultas (h1 y h2) con dos peso cada una (W7=0.7 y W8=0.8 para h1; W9=0.9 y W10=0.1 para h2) y dos nodos de salida (o1 y o2). El proceso normal para calcular el valor del nodo final se da, tanto con los nodos de entrada y los de capa oculta, mediante la sumatoria de producto de cada peso con su valor, es decir, mediante la fórmula de las RNA (colocar formula), al mismo tiempo que devuelve un valor del error cometido. Este último se calcula mediante la siguiente ecuación: Donde Tk es la salida correcta de cada nodo de salida, y Ok es la salida actual que cada uno genera. Con estos errores calculados, se retrocede hacia la capa oculta y se procede a calcular los nuevos pesos para sus nodos. La fórmula del cálculo de los mismos es:

  Donde Wjk representa el peso para cada nodo de la capa oculta, es decir, W7, W8, W9 y W10, los mismos que serán actualizados, L es el porcentaje de aprendizaje y Oj son los valores de estos dos nodos que entrarán a las salidas.

Estos nuevos pesos permitirán redefinir los errores de ambos nodos, con una pequeña diferencia en su cálculo:
El error de cada nodo de la capa oculta se obtiene multiplicando su valor por su complemento por la sumatoria del producto de sus pesos y los errores de los nodos de salida. Por ejemplo, para h1 sería (colocar fórmula).
Finalmente, se retrocede hacia los nodos de entrada y se repite el mismo proceso para la actualización de sus pesos y errores.

- **Función tangente hiperbólica**: Esta función está relacionada con una sigmoide bipolar. Sin embargo, sus salidas estarán en el rango de -1 y +1. Para redes neuronales, donde la velocidad es más importante que la forma de la función misma, es recomendable usar esta. Se representa como en la Figura 21 y su fórmula para calcular su nuevo valor es:
- **Función puramente lineal (purelin)**: Esta función se caracteriza porque su salida es igual a su entrada debido a su linealidad. Normalmente se usa para obtener los mismos valores de la entrada. Se representa como en la Figura 22 y su fórmula para calcular su nuevo valor es:
- **Función Unidad Lineal Rectificada (ReLU)**: Esta función se caracteriza por, además de conservar los valores positivos, convertir los valores negativos de entrada en 0, esto con la finalidad de no considerarlos en la siguiente capa de convolución como en el caso de procesamiento de imágenes (poner ref). Si bien tiene un buen desempeño en redes convolucionales y es muy usada para el procesamiento de imágenes, al no estar acotada pueden morirse demasiadas neuronas (poner ref). Se representa como en la Figura 23 y su fórmula para calcular su nuevo valor es:

Además de existir distintas funciones de activación, las redes neuronales artificiales se clasifican según la topología de red, siendo algunas de las más importantes (poner ref).

- **Red Neuronal Monocapa – Perceptrón simple**: Es la red neuronal más simple ya que está compuesta solamente de una capa de neuronas que componen varios nodos de entrada para proyectar una capa de neuronas de salida, como se aprecia en la Figura 24. Esta última capa se calcula usando la misma Ecuación 4 que implica la suma de productos de cada uno de los pesos de los nodos de entrada con sus instancias, añadiéndole finalmente el sesgo, aquel que controla la predisposición de la neurona a disparar un 1 o 0 independientemente de los pesos, para que el valor resultante se le aplique la función de activación que ayudarán a modelar funciones curvas o no triviales (poner ref).
- **Red Neuronal Multicapa – Perceptrón multicapa**: Con arquitectura similar al perceptrón simple, con el añadido de contener capas intermedias entre la capa de

neuronas de entrada y la de salida, conocidas como capas ocultas, como en el ejemplo de la Figura 25.

- **Redes Neuronales Convolucionales (CNN)**: También conocidas por su nombre en inglés Convolutional Neural Networks, se diferencia del perceptrón multicapa en que cada neurona no necesita estar unida con todas las que le siguen, sino más bien solo con un subgrupo de estas con el fin de reducir la cantidad de neuronas necesarias para su funcionamiento, como se observa en la Figura 26 (poner ref).

  Hoy en día, las redes neuronales convolucionales tienen múltiples usos desde que la idea fue concebida. Algunos de los problemas en las que pueden ser usados son de clasificación de objetos, recuperación de imágenes, detección y segmentación de objetos, distorsión y filtros de imágenes, por citar los ejemplos más comunes. El modelo de CNN más conocido es "AlexNet" (2012) por ser uno de los pioneros en clasificar imágenes (poner ref).

  Estas redes tienen su origen en el Neocognitron introducido por Fukushima en 1980 como modelo de red neuronal para el mecanismo de reconocimiento de patrón visual sin la enseñanza de un "profesor" (ver Figura 27), mismo que en el año 1998 sería mejorado por LeCun, Bottou, Bengio y Haffner al agregar un método de aprendizaje de gradiente aplicado al reconocimiento de documento basado en la propagación hacia atrás (ver Figura 28).

  Estos modelos se inspiraron en el estudio de la información visual en la corteza donde se ubican hasta 5 áreas. La primera, V1, contiene la información visual donde sus neuronas se ocupan de características visuales de bajo nivel, alimentando así a otras áreas adyacentes. Cada una de ellas se encarga de aspectos más específicos y detallados de la información obtenida. La idea de su implementación es la de solucionar el problema que surgen al escalar imágenes de mucha definición por las redes neuronales ordinarias. Por ello, este tipo de redes trabajan modelando de forma consecutiva piezas pequeñas de información para luego combinarlas en sus capas más profundas (poner ref).

  Su nombre deriva del concepto convolución. La convolución es un término en las matemáticas usado como operador matemático que convierte dos funciones f y g en una tercera función en donde la primera se superpone a una versión invertida y trasladada de la segunda, así como para denotar la distribución de la función de probabilidad de la suma de dos variables independientes aleatorias. Esta se da por la siguiente fórmula (poner ref).

  Donde el rango puede variar entre un conjunto finito (como en la fórmula desde 0 hasta un valor t) o uno infinito.

La estructura de las Redes Neuronales Convolucionales se constituye en tres tipos de capas (poner ref).

- **Capa convolucional (Convolutional Layer)**: Es la capa que hace distinta a esta red de otros tipos de redes neuronales artificiales. Se aplica la operación de la convolución, que recibe como entrada (input en inglés) a la imagen para luego aplicarle un filtro (kernel en inglés), devolviendo un mapa de las características de la imagen original, logrando así reducir el tamaño de los parámetros, como se observa en la Figura 29.

  Por ejemplo, en la anterior figura se tiene una imagen de entrada con dimensiones de 32 de alto, 32 de ancho y 3 de profundidad (32x32x3). A ella se le aplica un filtro de dimensiones (5x5x3) que recorrerá toda la imagen para extraer características de cada pixel. Tanto la profundidad de la entrada como del filtro siempre son iguales. El resultado de tomar un producto escalar entre el filtro y un pequeño fragmento de 5x5x3 de la imagen es un número, generando así un mapa de activación de nuevas dimensiones (28x28x1). Por cada n filtros aplicados a la entrada se generan n de estos mapas. Al final, la cantidad de mapas de activación determinará una nueva imagen de n de profundidad, como en la Figura 30.

  Asimismo, cada vez que se aplica una convolución a una imagen, se aplicará una función de activación como en la secuencia de la Figura 31.

  A nivel visual, en la Figura 32 se aprecia un ejemplo de los resultados de aplicar varias convoluciones a una imagen.

  Finalmente, se calcula el volumen de la dimensión de la salida de la Figura 33 mediante la siguiente ecuación: Se tiene una entrada de dimensiones (h x w x d). Se tiene un filtro de dimensiones (fh x fw x d).

- **Capa de reducción (Pooling Layer)**: Esta capa le sucede a la capa convolucional (luego de aplicar la función de activación). Sirve principalmente para reducir las dimensiones espaciales del volumen de la entrada (alto x ancho) para la siguiente capa convolucional. Sin embargo, no afecta la profundidad de la misma. Esta operación que realiza se le conoce también como "reducción de muestreo" debido a que, si bien logra reducir las dimensiones para procesar mejor en la siguiente capa, también conlleva perder información. Por el contrario de lo que se piensa, además de reducir la sobrecarga del cálculo para las siguientes capas, el modelo se beneficia también disminuyendo el sobreajuste.

  Para determinar las dimensiones de la nueva imagen generada (siempre que sea cuadrada, es decir, lados iguales como en la Figura 34) con esta capa, se aplica la siguiente fórmula:

Donde N es el tamaño del lado de la imagen de entrada, F es el tamaño del lado del filtro y Paso (Stride en inglés) es el número de desplazamiento de píxeles sobre la matriz de entrada. Por ejemplo, cuando el paso es 1, los filtros se mueven a 1 pixel por vez, cuando el paso es 2, se mueven a 2 píxeles (como en la Figura 35) y así sucesivamente (poner ref).

Si, por el contrario, se desea aplicar convolución a una imagen sin afectar sus dimensiones luego de pasar por la capa de reducción, se construye bordes de ceros de n píxeles. A este tamaño de borde se le llama Relleno (pad en inglés), por lo que el tamaño de la nueva salida se obtiene mediante la siguiente fórmula:

Existen diferentes tipos de reducción (poner ref):

- ◇ Max Pooling: Toma el elemento más grande dentro del mapa de características.
- ◇ Average Pooling: Toma el promedio de los elementos dentro del mapa de características.
- ◇ Sum Pooling: Toma la suma total de los elementos dentro del mapa de características.

- ○ **Capa totalmente conectada (Fully Connected Layer)**: Al final de las capas de convolución y reducción, se usan redes completamente conectadas a cada pixel considerando que cada uno como una neurona separada al igual que en una red neuronal regular (poner ref). En esta capa, se aplana la matriz de todas las características obtenidas anteriormente a un vector y se alinea en una capa completamente conectada a una red neuronal (Figura 36).

Para concluir, se muestra a continuación (Figura 37) la representación de la arquitectura completa de una Red Neuronal Convolucional resumiendo los conceptos anteriores.

- **Redes Neuronales Recurrentes (RNN)**: También conocidas por su nombre en inglés Recurrent Neural Networks, se caracterizan por no tener una estructura de capas como se aprecia en la Figura 38, sino más bien por permitir conexiones entre sus neuronas de manera arbitraria para crear temporalidad y que toda la red obtenga memoria. Todo esto permite generar una red muy potente para el análisis de secuencias, entre algunos ejemplos se mencionan el análisis de textos, sonidos o video (poner ref).

- **Máquina de Vectores de Soporte (SVM)**: Es un algoritmo usado para tareas de regresión y clasificación, buscando un hiperplano en un espacio N-dimensional que clasifique claramente los puntos de datos a partir de la distancia máxima entre los puntos de da-

tos de ambas clases. Para ello, maximiza la distancia del margen proporcionando cierto refuerzo para que los puntos de datos futuros puedan clasificarse con más confianza, es decir, que permita distinguir claramente dos clases, como se muestra en la Figura 39 (poner ref).

Este algoritmo tiene sus orígenes en la década de los años 60 en Rusia, desarrollados por Vapnik y Chervonenkis. Inicialmente se enfocó en el reconocimiento óptico de caracteres (OCR). Más tarde, los clasificadores de Vectores de Soporte se volvieron competitivos con los mejores sistemas disponibles en ese momento para resolver no solamente el anterior tipo de problema, sino también abarcar tareas de reconocimiento de objetos. En 1998, se publicó el primer manual de estos algoritmos por Burges. Y debido a sus grandes resultados obtenidos en la industria, actualmente se usa con frecuencia en el campo del aprendizaje automático (poner ref).

Los vectores de soporte hacen referencia a un pequeño subconjunto de las observaciones de entrenamiento que se utilizan como soporte para la ubicación óptima de la superficie de decisión (poner ref).

Una Máquina de Vectores de Soporte aprende la superficie decisión de dos clases distintas de los puntos de entrada. Como un clasificador de una sola clase, la descripción dada por los datos de los vectores de soporte es capaz de formar una frontera de decisión alrededor del dominio de los datos de aprendizaje con muy poco o ningún conocimiento de los datos fuera de esta frontera. Los datos son mapeados por medio de un kernel Gaussiano u otro tipo de kernel a un espacio de características en un espacio dimensional más alto, donde se busca la separación máxima entre clases. Cuando es traída de regreso al espacio de entrada, la función de frontera puede separar los datos en todas las clases distintas, cada una formando un agrupamiento. Esta teoría se basa en la idea de minimización de riesgo estructural (SRM), demostrando en muchas aplicaciones tener mejor desempeño que otros algoritmos de aprendizaje tradicional como las redes neuronales para resolver problemas de clasificación (poner ref).

Cabe mencionar que hay casos en que el conjunto de datos de dos clases puede ser separables no necesariamente de forma lineal. En las Figura 40 y Figura 41 se observan casos linealmente y no linealmente separables, respectivamente.

Lo que se debe hacer para el primer caso es crear el hiperplano a través de una función lineal w*z+b=0 y, definido el par (w,b), separar el punto xi según la función:

Para el segundo caso, debido a su mayor complejidad, se puede introducir algunas variables no-negativas a la función del hiperplano para hallar su valor óptimo; o también es viable utilizar una función kernel que calcule el producto punto de los puntos de entrada en el espacio de características Z, como se aprecia en la Figura 42.

- **Árboles de Decisión**: Representación visual de decisiones y toma de decisiones utilizada en la minería de datos para derivar una estrategia y alcanzar un objetivo particular. Se dibuja boca abajo con su raíz en la parte superior. Consta de nodos internos, los cuales se subdividen en ramas o bordes y su contenido, las hojas o decisiones (poner ref).

  Un árbol de decisión toma como entrada un objeto descrito a través de un conjunto de atributos y devuelve una "decisión". Estos pueden ser discretos o continuos. La salida puede tomar cualquiera de estos dos tipos de valores; en el caso que aprenda una función tomando valores discretos se le denominará clasificación, y en el caso que la función sea continua será llamada regresión. En las clasificaciones booleanas, es decir de dos valores o binaria, clasificará como verdadero (positivo) o falso (negativo). Para alcanzar una decisión, el árbol desarrolla una serie de pruebas a través de sus nodos y las ramas que salen del nodo son etiquetadas con los valores posibles de dicha propiedad. Además, cada nodo hojas del árbol representa el valor que ha de ser devuelto si es alcanzado (poner ref).

  Por ejemplo, representando un ejemplo de este algoritmo, se ilustra en la Figura 43 para decidir si se debe esperar por una mesa en un restaurante.

2.2.8. Natural Language Processing (NLP)

Naturalmano (Goyal y col., 2018). Otra definición para este término implica que es un campo especializado de la informática que es

De acuerdo con Goyal y col. (2018), e

2.3. Marco Conceptual

Para de

Capítulo 3

METODOLOGÍA DE LA INVESTIGACIÓN

3.1. Diseño de la investigación

En esta sección del documento se explicará cual es el diseño, el tipo y el enfoque del trabajo de investigación, así como también la población y la muestra.

3.1.1. Enfoque de la investigación

El presente trabajo tendrá un enfoque cuantitativo ya que se busca diseñar y desarrollar instrumentos, en este caso modelos predictivos, para responder al problema estudiado a partir de medición de datos históricos en la plataforma Kickstarter con herramientas basadas en la estadística y matemáticas que puedan ser interpretadas por cualquier investigador.

3.1.2. Alcance de la investigación

El alcance del presente trabajo será descriptivo ya que se recolectarán datos en un determinado rango de tiempo (desde 2009 hasta el presente año 2019) para describir el comportamiento de las campañas de proyectos tecnológicos en Kickstarter a partir de las características de sus variables y con ello, pronosticar su posible éxito o fracaso antes de finalizar la campaña con un nivel óptimo de precisión.

3.1.3. Tipo de la investigación

Para determinar el tipo de la investigación, primero es necesario definir el actual trabajo como Diseño Experimental ya que las variables que se tienen serán controladas, es decir, serán agregadas o quitadas en el o los modelos construidos en el experimento para analizar el impacto que este o estos tendrán en los resultados obtenidos. Dentro de esta categoría se clasifica como Diseño Experimental Puro ya que se busca medir la variable dependiente, en este caso Status (el estado actual del proyecto en Kickstarter) a partir de la manipulación de las demás variables independientes agregando o desagregándolas para comparar los rendimientos obtenidos de los instrumentos de medición y determinar cuáles de ellas finalmente serán tomadas en cuenta.

3.1.4. Descripción del prototipo de investigación

Teniendo como referencia y base principal el décimo antecedente explicado en el Capítulo II, la idea del prototipo final consistió en ensamblar las tres partes básicas de un proyecto: la primera consiste en el tratamiento de la metainformación (en la cual se realizarán, asimismo, tres experimentos independientes), el segundo, en el contenido visual y el último, el contenido textual respectivamente pero con el valor diferenciado de adaptar el modelo general de acuerdo a las variables y conjuntos de datos disponibles para el presente trabajo. Para ello, se representa cada una de las tres partes agrupadas en el marco de trabajo de la Figura 1.1.

3.2. Población y muestra

3.2.1. Población

La población que será considerada para el presente trabajo será de 27,251 proyectos en Kickstarter de la categoría tecnología de todas las subcategorías entre los periodos 2009-2019, en su mayoría del territorio de los Estados Unidos de América.

3.2.2. Muestra

Debido a que se 214 imágenes del contenido visual no pudieron re-dimensionarse, así como 2 proyectos no contaban con descripciones en el contenido textual, se procedió a remover los 216 proyectos incompletos tanto en la metainformación como en las otras bases de datos, resultando finalmente en 27,035 registros en cada una de los tres conjuntos de datos.

Sin embargo, la división en subconjuntos fue distinta en los tres casos y se dio de la siguiente manera:

- Para la metainformación y el contenido textual, el conjunto de datos total de cada uno fue dividido en un subconjunto de entrenamiento (80 %) y uno de prueba (20 %) siguiendo las proporciones dadas en el octavo antecedente.
- Para el contenido visual, el conjunto de datos total fue dividido en tres subconjuntos: entrenamiento (80 %), validación (10 %) y prueba (10 %) siguiendo las proporciones dadas en el décimo antecedente.

Nisi porta lorem mollis aliquam ut porttitor leo. Aenean pharetra magna ac placerat vestibulum. Est placerat in egestas erat imperdiet sed euismod. Velit euismod in pellentesque massa placerat. Enim praesent elementum facilisis leo vel fringilla. Ante in nibh mauris cursus mattis molestie a iaculis. Erat pellentesque adipiscing commodo elit at imperdiet dui accumsan sit. Porttitor lacus luctus accumsan tortor posuere ac ut. Tortor at auctor urna nunc id. A iaculis at erat pellentesque adipiscing commodo elit. La Figura 3.1 y el Cuadro 3.1

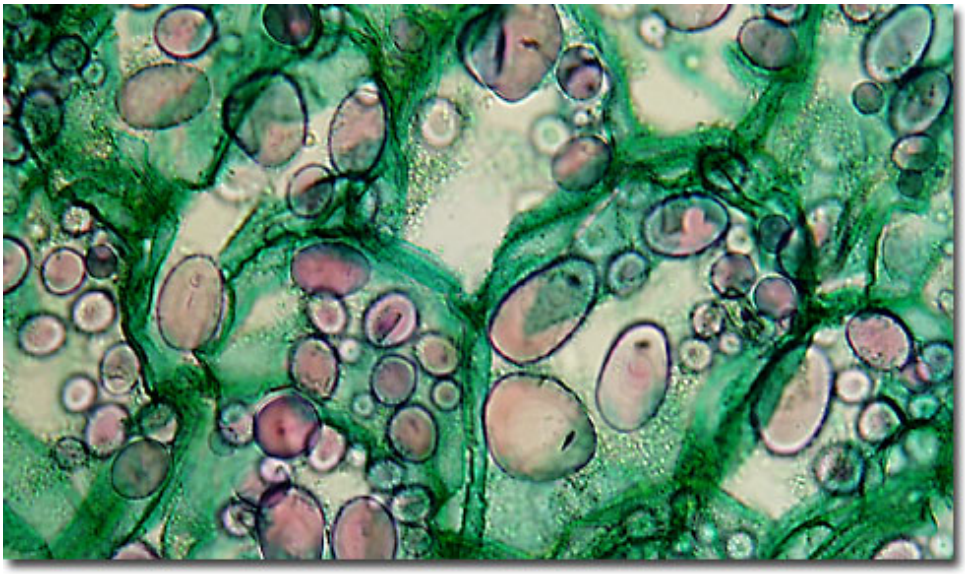

Figura 3.1: Prueba de Figura

3.2.3. Unidad de análisis

La unidad de análisis para el presente trabajo será un proyecto en Kickstarter de la categoría tecnología de cualquier subcategoría entre los periodos 2009-2019 dentro del territorio de los Estados Unidos de América.

3.3. Operacionalización de Variables

Nisi porta lorem mollis aliquam ut porttitor leo. Aenean pharetra magna ac placerat vestibulum. Est placerat in egestas erat imperdiet sed euismod. Velit euismod in pellentesque massa placerat. Enim praesent elementum facilisis leo vel fringilla. Ante in nibh mauris cursus mattis molestie a iaculis. Erat pellentesque adipiscing commodo elit at imperdiet dui accumsan sit. Porttitor lacus luctus accumsan tortor posuere ac ut. Tortor at auctor urna nunc id. A iaculis at erat pellentesque adipiscing commodo elit.

3.4. Instrumentos de medida

Nisi porta lorem mollis aliquam ut porttitor leo. Aenean pharetra magna ac placerat

- muscle and fat cells remove glucose from the blood,
- cells breakdown glucose via glycolysis and the citrate cycle, storing its energy in the form of ATP,
- liver and muscle store glucose as glycogen as a short-term energy reserve,
- adipose tissue stores glucose as fat for long-term energy reserve, and
- cells use glucose for protein synthesis.

3.5. Técnicas de recolección de datos

Nisi porta lorem mollis aliquam ut porttitor leo. Aenean pharetra magna ac placerat vestibulum. Est placerat in egestas erat imperdiet sed euismod. Velit euismod in pellentesque massa placerat. Enim praesent elementum facilisis leo vel fringilla. Ante in nibh mauris cursus mattis molestie a iaculis. Erat pellentesque adipiscing commodo elit at imperdiet dui accumsan sit. Porttitor lacus luctus accumsan tortor posuere ac ut. Tortor at auctor urna nunc id. A iaculis at erat pellentesque adipiscing commodo elit.

LaTeX is great at typesetting mathematics. Let $X_1, X_2, \ldots, X_n$ be a sequence of independent and identically distributed random variables with

$$S_n = \frac{X_1 + X_2 + \cdots + X_n}{n} = \frac{1}{n}\sum_{i}^{n} X_i \qquad \text{(Ecuación 3.1)}$$

La Ecuación Ecuación 3.1 denote their mean. Then as n approaches infinity, the random variables

$$\sqrt{n}(S_n - \mu)$$

converge in distribution to a normal $\mathcal{N}(0, \sigma^2)$.

3.6. Técnicas para el procesamiento y análisis de la información

Nisi porta lorem mollis aliquam ut porttitor leo. Aenean pharetra magna ac placerat vestibulum. Est placerat in egestas erat imperdiet sed euismod. Velit euismod in pellentesque massa placerat. Enim praesent elementum facilisis leo vel fringilla. Ante in nibh mauris cursus mattis molestie a iaculis. Erat pellentesque adipiscing commodo elit at imperdiet dui accumsan sit. Porttitor lacus luctus accumsan tortor posuere ac ut. Tortor at auctor urna nunc id. A iaculis at erat pellentesque adipiscing commodo elit.

You can make lists with automatic numbering ...

1. Like this,
2. and like this.

... or bullet points ...

- Like this,
- and like this.

3.7. Cronograma de actividades y presupuesto

Nisi porta lorem mollis aliquam ut porttitor leo. Aenean pharetra magna ac placerat vestibulum. Est placerat in egestas erat imperdiet sed euismod. Velit euismod in pellentesque massa placerat. Enim praesent elementum facilisis leo vel fringilla. Ante in nibh mauris cursus mattis molestie a iaculis. Erat pellentesque adipiscing commodo elit at imperdiet dui accumsan sit. Porttitor lacus luctus accumsan tortor posuere ac ut. Tortor at auctor urna nunc id. A iaculis at erat pellentesque adipiscing commodo elit.

Item	Quantity
Widgets	42
Gadgets	13

Tabla 3.1: An example table.

Capítulo 4

DESARROLLO DEL EXPERIMENTO

4.1. X

Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like “Huardest gefburn? Kjift ”not at all!...

4.2. Y

Nisi porta lorem mollis aliquam ut porttitor leo. Aenean pharetra magna ac placerat vestibulum. Est placerat in egestas erat imperdiet sed euismod. Velit euismod in pellentesque massa placerat. Enim praesent elementum facilisis leo vel fringilla. Ante in nibh mauris cursus mattis molestie a iaculis. Erat pellentesque adipiscing commodo elit at imperdiet dui accumsan sit. Porttitor lacus luctus accumsan tortor posuere ac ut. Tortor at auctor urna nunc id. A iaculis at erat pellentesque adipiscing commodo elit.

Item	Quantity
Widgets	42
Gadgets	13

Tabla 4.1: An example table.

4.3. Z

Nisi porta lorem mollis aliquam ut porttitor leo. Aenean pharetra magna ac placerat vestibulum. Est placerat in egestas erat imperdiet sed euismod. Velit euismod in pellentesque massa placerat. Enim praesent elementum facilisis leo vel fringilla. Ante in nibh mauris cursus mattis molestie a iaculis. Erat pellentesque adipiscing commodo elit at imperdiet dui accumsan sit. Porttitor lacus luctus accumsan tortor posuere ac ut. Tortor at auctor urna nunc id. A iaculis at erat pellentesque adipiscing commodo elit.

El paper es citado y el otro paper .

Capítulo 5

ANÁLISIS Y DISCUSIÓN DE RESULTADOS

5.1. X

Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like “Huardest gefburn? Kjift ”not at all!...

5.2. Y

Nisi porta lorem mollis aliquam ut porttitor leo. Aenean pharetra magna ac placerat vestibulum. Est placerat in egestas erat imperdiet sed euismod. Velit euismod in pellentesque massa placerat. Enim praesent elementum facilisis leo vel fringilla. Ante in nibh mauris cursus mattis molestie a iaculis. Erat pellentesque adipiscing commodo elit at imperdiet dui accumsan sit. Porttitor lacus luctus accumsan tortor posuere ac ut. Tortor at auctor urna nunc id. A iaculis at erat pellentesque adipiscing commodo elit.

Item	Quantity
Widgets	42
Gadgets	13

Tabla 5.1: An example table.

5.3. Z

Nisi porta lorem mollis aliquam ut porttitor leo. Aenean pharetra magna ac placerat vestibulum. Est placerat in egestas erat imperdiet sed euismod. Velit euismod in pellentesque massa placerat. Enim praesent elementum facilisis leo vel fringilla. Ante in nibh mauris cursus mattis molestie a iaculis. Erat pellentesque adipiscing commodo elit at imperdiet dui accumsan sit. Porttitor lacus luctus accumsan tortor posuere ac ut. Tortor at auctor urna nunc id. A iaculis at erat pellentesque adipiscing commodo elit.

Capítulo 6

CONCLUSIONES Y RECOMENDACIONES

6.1. Conclusiones

Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like “Huardest gefburn? Kjift ”not at all!...

6.2. Recomendaciones

Nisi porta lorem mollis aliquam ut porttitor leo. Aenean pharetra magna ac placerat vestibulum. Est placerat in egestas erat imperdiet sed euismod. Velit euismod in pellentesque massa placerat. Enim praesent elementum facilisis leo vel fringilla. Ante in nibh mauris cursus mattis molestie a iaculis. Erat pellentesque adipiscing commodo elit at imperdiet dui accumsan sit. Porttitor lacus luctus accumsan tortor posuere ac ut. Tortor at auctor urna nunc id. A iaculis at erat pellentesque adipiscing commodo elit.

Anexos

Anexos A

Anexo I: Matriz de Consistencia

PROBLEMAS	OBJETIVOS	HIPÓTESIS
Problema General	Objetivo General	Hipótesis General
Bajos niveles de precisión de modelos entrenados de Aprendizaje Automático para cualquier categoría para predecir estado de financiamiento de proyectos de tecnología.	Construir modelo(s) de Aprendizaje Automático entrenado(s) para predecir correctamente proyectos de tecnología con nivel de precisión aceptable.	El modelo entrenado de Aprendizaje Automático logrará predecir correctamente proyectos de tecnología con nivel de precisión aceptable.
Problemas Específicos	Objetivos Específicos	Hipótesis Específicas
Variables de proyectos no normalizadas y varianzas altas.	Normalizar variables de proyectos y reducir niveles altos de varianza.	Las variables de los proyectos descargados se normalizarán y se reducirán los niveles altos de varianza.
Datos faltantes o incompletos de proyectos.	Eliminar datos faltantes o incompletos de proyectos.	Los datos faltantes o incompletos de los proyectos serán eliminados.
Parámetros de modelos no ajustados.	Ajustar parámetros de modelos.	Los parámetros de los modelos usados serán ajustados.
Sobreajuste de aprendizaje de modelos y clasificación incorrecta de las dos clases del estado final de financiamiento (exitoso o fracasado).	Evitar sobreajuste de aprendizaje de modelos.	Se evitará el sobreajuste de aprendizaje de modelos para clasificar correctamente las dos clases del estado final de financiamiento.
Predicción incorrecta de estado de financiamiento de un proyecto tecnológico.	Predecir correctamente el estado final de financiamiento de cualquier proyecto tecnológico (éxito o fracaso).	El estado final de financiamiento de cualquier proyecto tecnológico será predicho correctamente.

Tabla A.1: Matriz de consistencia. Fuente: Elaboración propia

Anexos B

Anexo II: Resumen de Papers investigados

Tipo	N°	Título	Autor	Año	País	Fuente
Problema	1	Copper price estimation using bat algorithm	Dehghani Bogdanovic	2018	United Kingdom	Resources Policy
	2	Alternative techniques for forecasting mineral commodity prices	Cortez, Saydam, Coulton, Sammut	2018	Netherlands	International Journal of Mining Science and Technology
Propuesta	3	Prediction of the crude oil price thanks to natural language processing applied to newspapers	Trastour, Genin, Morlot	2016	USA	Standfort University ML repository
	4	Stock Price Prediction Using Deep Learning	Tipirisetty	2018	USA	Master's Theses San Jose State University
	5	Deep Learning for Stock Prediction Using Numerical and Textual Information	Akita, R., Yoshihara, A., Matsubara, T., Uehara, K.	2016	USA	2016 IEEE/ACIS 15th International Conference on Computer and Information Science (ICIS)
Técnica	6	Stock Prices Prediction using the Title of Newspaper Articles with Korean Natural Language Processing	Yun, Sim, Seok	2019	Japan	2019 International Conference on Artificial Intelligence in Information and Communication (ICAIIC)
	7	A Method of Optimizing LDA Result Purity Based on Semantic Similarity	Jingrui, Z., Qinglin, W., Yu, L., Yuan, L.	2017	China	2017 32nd Youth Academic Annual Conference of Chinese Association of Automation (YAC)
	8	Qualitative Stock Market Predicting with Common Knowledge Based Nature Language Processing: A Unified View and Procedure	Rao, D., Deng, F., Jiang, Z., Zhao, G.	2015	USA	2015 7th International Conference on Intelligent Human-Machine Systems and Cybernetics
	9	Fuzzy Bag-of-Words Model for Document Representation	Zhao, R., Mao, K.	2018	USA	IEEE Transactions on Fuzzy Systems (Volume: 26 , Issue: 2 , April 2018)

Tabla B.1: Cuadro Resumen de Papers investigados. Fuente: Elaboración propia

BIBLIOGRAFÍA

Alpaydin, E. (2014). *Introduction to Machine Learning* (Tercera edición). MIT Press.

Asociación de Emprendedores de Perú. (2018). Avances y limitaciones del emprendimiento peruano. https://asep.pe/index.php/avances-limitaciones-emprendimiento-peruano/

Beckwith, J. (2016). Predicting Success in Equity Crowdfunding. *Joseph Wharton Scholars*. http://repository.upenn.edu/joseph_wharton_scholars/25

Chen, S.-Y., Chen, C.-N., Chen, Y.-R., Yang, C.-W. & Lin, W.-C. (2015). Will Your Project Get the Green Light? Predicting the Success of Crowdfunding Campaigns. http://aisel.aisnet.org/pacis2015/79

Cheng, C., Tan, F., Hou, X. & Wei, Z. (2019). Success Prediction on Crowdfunding with Multimodal Deep Learning, 2158-2164. https://www.ijcai.org/proceedings/2019/0299.pdf

Gartner. (2019). Gartner IT Glossary - Machine Learning. https://www.gartner.com/it-glossary/machine-learning/

Goyal, P., Pandey, S. & Jain, K. (2018). Deep learning for natural language processing. *Deep Learning for Natural Language Processing: Creating Neural Networks with Python [Berkeley, CA]: Apress*, 138-143.

Jin, B., Zhao, H., Chen, E., Liu, Q. & Ge, Y. (2019). Estimating the Days to Success of Campaigns in Crowdfunding: A Deep Survival Perspective. http://staff.ustc.edu.cn/~cheneh/paper_pdf/2019/Binbin-Jin-AAAI.pdf

Kamath, R. S. & Kamat, R. K. (2018). Supervised Learning Model For Kickstarter Campaigns With R Mining. *International Journal of Information Technology, Modeling and Computing (IJITMC)*, *4*(1). https://doi.org/10.5281/zenodo.1228716

Kaur, H. & Gera, J. (2017). Effect of Social Media Connectivity on Success of Crowdfunding Campaigns. *Procedia Computer Science*, *122*, 767-774. https://doi.org/10.1016/j.procs.2017.11.435

Kickstarter. (s.f.-a). *Acerca de nosotros: Kickstarter*. https://www.kickstarter.com/about?ref=global-footer

Kickstarter. (s.f.-b). *Financiamiento: Kickstarter*. https://www.kickstarter.com/help/handbook/funding?lang=es

Kickstarter. (s.f.-c). *Prensa: Kickstarter*. https://www.kickstarter.com/press?ref=hello

Li, Y., Rakesh, V. & Reddy, C. K. (2016). Project Success Prediction in Crowdfunding Environments, 247-256. https://doi.org/10.1145/2835776.2835791

Microsoft. (2018). Algoritmos de minería de datos (Analysis Services: Minería de datos). https://docs.microsoft.com/es-mx/sql/analysis-services/data-mining/data-mining-algorithms-analysis-services-data-mining?view=sql-server-2017

Poole, D., Mackworth, A. & Goebel, R. (1998). Computational Intelligence: A Logical Approach. *Oxford University Press*.

Redacción Gestión. (2015). Emprendimiento en el Perú se origina más por oportunidades de negocio que por desempleo. *Diario Gestión*. https://gestion.pe/economia/emprendimiento-peru-origina-oportunidad-negocio-desempleo-80578

Redacción Gestión. (2018). Perú es el tercer país con mayor cantidad de emprendimientos en fase temprana a nivel mundial. *Diario Gestión*. https://gestion.pe/economia/peru-tercer-pais-mayor-cantidad-emprendimientos-fase-temprana-nivel-mundial-240264

Russell, S. & Norvig, P. (2004). *Inteligencia Artificial: Un Enfoque Moderno* (J. M. Corchado Rodríguez, F. Martín Rubio, J. M. Cadenas Figueredo, L. D. Hernández Molinero, E. Paniagua Arís, R. Fuentetaja Pinzán, M. Robledo de los Santos & R. Rizo Aldeguer, Trad.; Segunda edición). Pearson Educación, S.A. https://luismejias21.files.wordpress.com/2017/09/inteligencia-artificial-un-enfoque-moderno-stuart-j-russell.pdf

Russell, S. & Norvig, P. (2009). *Inteligencia Artificial: Un Enfoque Moderno* (Tercera edición). Prentice Hall.

Sancho Caparrini, F. (2018). *Clasificación Supervisada y No Supervisada* (Publicación). Universidad de Sevilla. location. http://www.cs.us.es/~fsancho/?e=77

Sandoval, L. (s.f.). Barreras del Emprendedor ¿Por qué cuesta tanto hacerlo? https://www.emprender-facil.com/es/barreras-del-emprendedor/

Solidaridad Latina. (s.f.). ¿Cómo funciona el crowdfunding en Latinoamérica? https://solidaridadlatina.com/actualizacion/como-funciona-crowdfunding-latinoamerica/

The Hustle. (2019). What are your chances of successfully raising money on Kickstarter? https://thehustle.co/crowdfunding-success-rate

Universo Crowdfunding. (s.f.). ¿Qué es el crowdfunding? https://www.universocrowdfunding.com/que-es-el-crowdfunding/

Yu, P.-F., Huang, F.-M., Yang, C., Liu, Y.-H., Li, Z.-Y. & Tsai, C.-H. (2018). Prediction of Crowdfunding Project Success with Deep Learning, 1-8. https://doi.org/10.1109/ICEBE.2018.00012

Yuan, H., Lau, R. Y. & Xu, W. (2016). The Determinants of Crowdfunding Success: A Semantic Text Analytics Approach. *Decision Support Systems*, *91*, 67-76. https://doi.org/10.1016/j.dss.2016.08.001

Zambrano, J. (2018). ¿Aprendizaje supervisado o no supervisado? Conoce sus diferencias dentro del machine learning y la automatización inteligente. *Medium*. https://medium.com/@juanzambrano/aprendizaje-supervisado-o-no-supervisado-39ccf1fd6e7b

Zhou, M., Zhang, X., Wang, A. G., Du, Q., Qiao, Z. & Fan, W. (2018). Money Talks: A Predictive Model on Crowdfunding Success Using Project Description. *20*(2), 259-274. https://doi.org/10.1007/s10796-016-9723-1