# Predicción de la velocidad de adopción de perros y gatos a partir de modelos clásicos de machine learning y deep learning con imágenes y textos

María Fernanda Giraldo Plaza

Claudia Yaneth Valencia Morales

Monografía presentada para optar al título de Especialista en Analítica y Ciencia de Datos

Asesor

Antonio Jesús Tamayo Herrera, Doctor (PhD) Ciencias de la Computación

Universidad de Antioquia
Facultad de Ingeniería

Especialización en Analítica y Ciencia de Datos

Medellín, Antioquia, Colombia

2021

| Cita | Giraldo Plaza y Valencia Morales [1] |
|---|---|
| **Referencia**<br><br>Estilo IEEE (2020) | [1] M. Giraldo Plaza y C. Valencia Morales, “Predicción de la velocidad de adopción de perros y gatos a partir de modelos clásicos de machine learning y deep learning con imágenes y textos”, Trabajo de grado especialización, Especialización en Analítica y Ciencia de Datos, Universidad de Antioquia, Medellín, Antioquia, Colombia, 2021. |




Especialización en Analítica y Ciencia de Datos, Cohorte II.

# TABLA DE CONTENIDO

# LISTA DE TABLAS

## LISTA DE FIGURAS

## I. RESUMEN

Muchos animales en situación de calle sufren y son sacrificados en refugios todos los días en todo el mundo por falta de recursos económicos y financieros para el sostenimiento y su debido cuidado.

El objetivo de este trabajo consiste en utilizar los datos de una competencia de *Kaggle* donde la finalidad de esta es predecir la velocidad de adopción de dos tipos de animales (perro y gato). Los datos son obtenidos de una base de *Petfinder* de Malasia, con capacidad de búsqueda de animales que necesitan un hogar. Esta base es un directorio de casi 150,000 animales. Las tasas de adopción de animales se encuentran correlacionadas con los metadatos asociados a sus perfiles en línea como el texto descriptivo, las características que se evidencian en las fotografías y las variables que se consideran importantes predictores de adopción como la edad, el sexo, la raza, entre otros.

Para desarrollar el propósito definido se ejecutan modelos de aprendizaje automático (*machine learning*) y aprendizaje profundo (*deep learning*) con el fin de implementar algoritmos de predicción de la velocidad de adopción de mascotas según las características de texto; imágenes y atributos, los resultados obtenidos indican que esta es una base de datos compleja ya que los modelos implementados a partir de los datos tienen un mal rendimiento para la clasificación planteada.

***Palabras claves* — Velocidad de adopción, machine learning, deep learning, predicción, datos, mascotas.**

## II. ABSTRACT

Every day many stray animals suffer and are sacrificed in shelters around the world due to a lack of economic and financial resources for their maintenance and proper care.

The objective of this study is to use the data from a Kaggle competence, where the purpose was to predict the adoption rate for two types of animals (dogs and cats). The data was obtained through the Petfinder database from Malaysia, which has the capacity to search for animals that need a home. This database is a directory of around 150,000 animals. The adoption rate of the animals is correlated with the metadata associated with the online profiles as a descriptive text. The characteristics that are evident in the pictures and the variables that are considered important by the adoption predictors such as age, sex, breeds, among others.

To develop the defined purpose, machine learning, and deep learning models were run in order to implement algorithms to predict the speed of adoption based on the text characteristics, images, and attributes. The results obtained indicate that this is a complex database since the models implemented from the data have a poor performance for the proposed classification.

***Keywords*** — **Adoption speed, machine learning, deep learning, prediction, data, pets.**

# III. INTRODUCCIÓN

Existen en el mundo actual una serie de situaciones críticas relacionadas con los animales en situación de calle, puesto que, dadas las limitaciones presupuestales de los gobiernos y las restricciones financieras de los hogares, la adopción se convierte en un mecanismo que busca cerrar esa brecha y mejorar las condiciones de salud públicas relacionada con la tenencia de mascotas.

Debido a la problemática descrita anteriormente, se pueden salvar muchas vidas si se les encuentra un hogar. Por esta razón, se eligió una competencia de *Kaggle* de la empresa *PetFinder.my* [1], la cual cuenta con una plataforma de bienestar animal líder en Malasia desde 2008 que tiene una base de datos de más de 150.000 animales.

Principalmente la organización menciona que las tasas de adopción de animales están fuertemente correlacionadas con los metadatos asociados con sus perfiles en línea, como el texto descriptivo, los datos tabulares y las características de las fotografías. Con estos datos se busca implementar algoritmos para predecir la adopción de las mascotas, específicamente, qué tan rápido se adopta una mascota después de que ingresa a la base de datos.

Cabe resaltar, que la motivación principal para realizar este trabajo fue la cantidad de información y el reto de analizar diversas fuentes de datos para lograr una predicción del tiempo de adopción de las mascotas dependiendo de su perfil y sus características. Dicha predicción se realiza a través de la combinación de algoritmos utilizados de aprendizaje automático clásico y aprendizaje profundo, los cuales permiten hacer un análisis de los datos para llevar a cabo la tarea establecida y así determinar que a través de la identificación de ciertas particularidades se puede mejorar la velocidad de adopción.

En este documento se podrá observar la definición del problema donde se plantea la pregunta a resolver, los objetivos del trabajo, el marco teórico que comprende la generalidad de los conceptos de inteligencia artificial y los trabajos relacionados que implementan modelos para resolver situaciones de adopción, retorno al refugio y sacrificio; la metodología que se desarrolla, los resultados y la discusión de estos y finalmente las conclusiones.

## IV. PLANTEAMIENTO DEL PROBLEMA

Malasia está situada en Asia sudoriental, su capital es *Kuala Lumpur*. Este país cuenta con una población de 32.939.000 habitantes, en su mayoría las personas profesan el islam y su régimen político es la monarquía electiva constitucional [2].

Según Cesce [3], el principal grupo étnico en Malasia es el malayo (*bumiputras*), que comprende aproximadamente el 50% de los casi 33 millones de habitantes del país y que son mayoritariamente de religión musulmana. Le siguen en importancia la minoría china (23% de la población) y la hindú (7%). También existen otros grupos étnicos autóctonos minoritarios, que en conjunto representan cerca del 12%, y una notable comunidad foránea que supone casi el 8%. Es importante resaltar que las desigualdades en la distribución de la riqueza en Malasia son muy inferiores y cuenta con una renta media alta.

En Malasia se creó una ley de bienestar animal en el año 2015 (*Animal Welfare Act 2015 - Act 772*) con el objetivo de fomentar el bienestar y la propiedad responsable de los animales, y sancionar la crueldad animal. Para el cumplimiento de esta ley se asignaron 400 agentes de bienestar animal encargados de hacer cumplir la ley, supervisando el trabajo de las asociaciones de protección animal [4].

Según esta ley se requiere licencias para todas las personas y empresas que disponen de animales, prohíbe la cría de animales para la investigación o la enseñanza, prohíbe disparar contra perros callejeros y asigna una multa por crueldad animal y puede imponer penas privativas de la libertad por meses o años de prisión, dependiendo del caso [5].

Ahora bien, no es suficiente que exista una normatividad que proteja a los animales en situación de calle, es necesario generar condiciones apropiadas para mejorar su calidad de vida.

Con relación a lo anterior, Wong Noel [6] señala que en Malasia a pesar de que ha aumentado el número de animales rescatados de las calles, no hay una correlación positiva con el número de adopciones, lo cual genera un alto hacinamiento y una desmejora en las condiciones de salud y bienestar de las mascotas en los refugios dispuestos como albergues temporales.

Así mismo, cita que *My Forever Doggo (MFD)*, es una iniciativa privada que busca conectar a perros ubicados en refugios con posibles adoptantes utilizando las redes sociales.

Este colectivo logró determinar que la falta de visibilidad pública de un refugio, el bajo tráfico peatonal, la poca iluminación que tiene la ubicación y las experiencias de refugio desfavorables habían contribuido a las bajas tasas de adopción en *Klang Valley*, un conglomerado urbano en *Kuala Lumpur*.

De igual manera, es posible afirmar que otro de los problemas evidenciados alrededor de la adopción de mascotas es el tiempo que transcurre entre tener una mascota disponible y asignar un dueño o familia que se responsabilice de su cuidado, y adicionalmente que la elección del perro o gato cumpla con las expectativas del adoptante, con el fin de evitar el abandono de la mascota. Además, es importante anotar que las organizaciones dedicadas a esta labor no tienen subvenciones en este país por parte del estado, por lo tanto, tienen limitaciones de recursos financieros y de infraestructura para el sostenimiento de los animales a mediano y largo plazo.

Por este motivo *PetFinder* ha sido una plataforma de bienestar animal líder en Malasia desde 2008 [1], *PetFinder* colabora estrechamente con los amantes de los animales, los medios de comunicación, las corporaciones y las organizaciones globales para mejorar el bienestar animal. Esta plataforma pone a disposición información que permite predecir la velocidad en la que se adopta una mascota, considerando esta como el tiempo que transcurre entre la creación del perfil de la mascota o grupo de mascotas y la adopción de éstas. La información disponible corresponde a características del perfil que indica información como tipo de animal, género, raza, entre otras; también se cuenta con la descripción del perfil e imágenes asociadas a este.

De acuerdo con lo anterior, se plantea la siguiente pregunta a resolver: ¿Cómo implementar un modelo de predicción que permita identificar el tiempo en el cual será adoptada una mascota según determinadas características?

# V. OBJETIVOS

## *A. OBJETIVO GENERAL*

Implementar diferentes modelos predictivos de aprendizaje automático clásico y aprendizaje profundo para la determinación del tiempo de adopción de perros y gatos.

## *B. OBJETIVOS ESPECÍFICOS*

Obtener una base de datos de entrenamiento etiquetada que contenga el tiempo de adopción de perros y gatos.

Realizar un análisis descriptivo de las variables cuantitativas y cualitativas de las mascotas.

Implementar algoritmos de aprendizaje automático y aprendizaje profundo para la predicción de la velocidad de adopción.

Reducir la dimensión por selección de características y aplicar reducción de dimensión por análisis de componentes principales (*Principal Component Analysis - PCA*).

## VI. MARCO TEÓRICO

En esta sección del trabajo se abordará la búsqueda realizada en torno a las herramientas utilizadas en el campo de los algoritmos de aprendizaje automático, aprendizaje profundo y los trabajos relacionados que evalúan la adopción de animales con la implementación de dichos algoritmos.

Sarker [7] hace referencia a que el mundo en la era de la cuarta revolución industrial cuenta con una gran cantidad de datos que permiten generar algoritmos para el análisis de estos a gran escala. A partir de esta disponibilidad de los datos se han venido realizando clasificaciones de la información, como lo son los datos estructurados, semi estructurados y no estructurados. De igual forma, los algoritmos para procesar y aprender de dicha información también tienen su distribución: los cuales son supervisados, no supervisados y semi supervisados.

Los tipos de datos estructurados son aquellos que regularmente se encuentran en formato tabular que son de fácil acceso y manipulación; los datos no estructurados son complejos en términos de captura, procesamiento y análisis, a este corresponde el material de imágenes, videos, texto, entre otros; los datos semiestructurados no se almacenan en una base de datos relacional, pero sí poseen propiedades organizativas que permiten su análisis, el ejemplo de estos son archivos *HTML*, *XML* y *JSON*, hay una última clasificación que corresponde a los metadatos que hace referencia a información relevante de los datos.

La literatura indica que existen varios tipos de algoritmos de aprendizaje para gestionar los tipos de datos que son mencionados anteriormente, dentro de estos se encuentran los supervisados, estos se refieren a aquellos que aprenden de los datos etiquetados. No supervisados, que corresponde a los algoritmos para gestionar datos no etiquetados. Semi supervisados, que abarcan de manera combinada los datos etiquetados y no etiquetados; y por último se tienen los algoritmos por refuerzo que se basa en recompensas y penalizaciones. Para cada una de estas clasificaciones se usan algoritmos específicos que permiten solucionar problemas de predicción, interacción con objetos, recomendación de productos en línea considerando el perfil que se construye a partir de los datos de un usuario, procesamiento del lenguaje natural, análisis de sentimientos y reconocimiento de imágenes.

Si bien es importante conocer las herramientas que son utilizadas para el procesamiento de los datos, es necesario comprender conceptos como la minería de datos el cual es utilizado para denotar la extracción de información, de los datos que puede ser útil para algún proceso. Abarca un conjunto de técnicas enfocadas en la extracción de conocimiento implícito en las bases de datos.

Un proceso típico de minería de datos considera la selección de un conjunto de datos, que valora las variables dependientes y las objetivo; el análisis de las propiedades de los datos a través de estadística descriptiva; la transformación del conjunto de datos de entrada, donde el objetivo es adaptar la técnica de minería de datos que mejor se acople al problema; seleccionar y aplicar la técnica de minería de datos y finalmente evaluar los resultados obtenidos. Estas técnicas provienen de la inteligencia artificial y la estadística, en la fase de minería de datos se decide cual es la tarea a realizar, ya sea clasificar o predecir.

Las técnicas descriptivas corresponden al *clustering* y segmentación, escalamiento de los datos, reglas de asociación y dependencia, análisis exploratorio y reducción de la dimensión; las técnicas de predicción son regresión y series temporales, análisis discriminante, métodos bayesianos, algoritmos genéticos, árboles de decisión y redes neuronales [8].

Se hace referencia de igual forma a la comprensión de las técnicas de aprendizaje profundo para la segmentación de imágenes; las revisiones detallan la implementación de estos modelos, describiendo las técnicas disponibles para abordar este tipo de problemas. Se tienen las redes neuronales convolucionales que permiten la salida lineal que se requiere en una clasificación; para convertir volúmenes de activaciones de dos dimensiones se requiere aplanar la data de las imágenes lo cual permite la ejecución de redes completamente conectadas para obtener la distribución de probabilidad que indica a qué categoría corresponde la imagen que se estudia. Otra técnica utilizada corresponde a *DeepMask* y *SharpMask* donde el modelo es capaz de realizar múltiples tareas, un modelo crea una clasificación a nivel de píxel y la segunda rama genera una puntuación correspondiente a la detección de objetos. Otra rama similar que empezó a desarrollarse es la localización de objetos, la tarea para estos modelos es localizar objetos específicos en las imágenes estudiadas. Otras técnicas utilizadas en aprendizaje profundo son los modelos adversarios en los cuales el aprendizaje consta de dos redes: una red generativa y otra discriminante; la

generativa se encarga de producir imágenes como las del conjunto de entrenamiento, utilizando una distribución que permite ingresar ruido al modelo, la red discriminante se encarga de decidir si la entrada corresponde a un dato de entrenamiento o un dato originado por la red generativa. Los modelos secuenciales segmentan el objeto a la salida de la red, las principales arquitecturas utilizadas en este tipo de problemas corresponden a *LSTM* convolucionales, redes recurrentes, entre otras [9].

Es importante destacar que las técnicas anteriormente mencionadas se implementan en diversos sectores de la industria con el fin de optimizar los procesos, mejorar las ventas, reducir costos, entre otros; de igual manera se tienen implementaciones de carácter social que permite solucionar problemas con el fin de efectuar aportes al bienestar de la comunidad. De acuerdo con el problema definido se realiza la búsqueda de trabajos relacionados con implementaciones de modelos de aprendizaje automático y aprendizaje profundo que aborden soluciones al problema.

Bradley et al. [10] indican que alrededor de 6 a 8 millones de animales ingresan a refugios por año y aproximadamente de 3 a 4 millones (50%) son sacrificados y entre el 10% - 25% son sacrificados directamente por el refugio dado el hacinamiento que se genera en este. Se encuentran trabajos que, a partir de una regresión logística, una red neuronal artificial, el aumento del gradiente y los algoritmos de árboles de decisión aleatorios predicen la duración en el refugio de los animales. Estos modelos son evaluados a partir de tres métricas: la precisión, recuperación y puntuación.

De acuerdo con lo anterior, el problema trata de poder identificar los mejores modelos partiendo de la necesidad de predecir la eutanasia, la adopción y el retorno al refugio, los mejores modelos para este estudio son el *gradient boosting* y *random forest*. Este estudio también proporciona información asociada a los factores que influyen de manera significativa en la duración de la mascota en el albergue, estos indican que la edad, la raza, el pelaje y el color tienen un impacto importante en la duración del animal en el refugio.

Adicionalmente, se encuentra un trabajo donde se realiza un piloto durante 8 semanas con 55 perros, el cual mediante el uso de monitores cuanti-métricos en los perros adoptados busca obtener datos de los adoptantes por medio de una aplicación de teléfono inteligente [11]. El objetivo de esta investigación era reducir las devoluciones de los animales y

aumentar la unión entre los perros adoptados de *Humane Society of Silicon Valley* (*HSSV*) y sus adoptantes.

Según Alcaidinho et al.[11] durante estas 8 semanas, 18 perros en el grupo del experimento fueron adoptados. Después de realizar encuestas a 12 adoptantes de los 18 y obtener estadísticas, el 40.7% de los perros adoptados fueron devueltos en el plazo de una semana. El estudio también arrojó que después de una semana el promedio de devolución fue de 33.8 días, la justificación principal de la devolución se debió a problemas de comportamiento del animal, sin embargo, éste también indicó que conocer la actividad del perro a lo largo del día, incluso cuando el adoptante no estaba cerca, podría aumentar el vínculo entre el posible adoptante y el perro. Además, los encuestados indicaron que el uso de la aplicación les ayudó a satisfacer mejor las necesidades de actividad de sus perros y aumentó el vínculo entre ellos. Esta herramienta es de mucha utilidad porque permite a los adoptantes y al albergue prevenir las renuncias y por ende el retorno de animales.

Finalmente, se observan los resultados obtenidos por Zhang et al. [12]al predecir la velocidad de adopción de mascotas mediante una base de datos de *PetFinder* de *Kaggle*, que considera tres tipos de datos; tabulares relacionados en un CSV donde se detalla información del animal como lo es el tipo (perro o gato), raza, género, color, pelaje, entre otras, con un total de 14.993 registros de perfiles de animales y 23 características; archivos tipo JSON donde se encuentra la descripción del animal y finalmente imágenes asociadas al perfil de los animales. En este trabajo se realiza preprocesamiento de los datos mediante el *one hot encoding*, considerando que se tratan de variables categóricas. Realizan el entrenamiento y predicción mediante dos tipos de modelos: modelos clásicos de *Machine Learning* y modelos de *Deep Learning*. Para los modelos de *Machine Learning* consideran algoritmos como *LogisticRegression*, *Naive Bayes*, *SVM*, *Decision Tree*, *Random Forest* y *Gradient Boosting*, donde la mejor predicción la entrega el *Random Forest* con una precisión del 38%; para los modelos de *Deep Learning* se consideran el *Fully connected*, *LSTM* y *combined*, la mejor predicción la entrega el modelo *Fully connected* con una precisión del 39%.

## VII. METODOLOGÍA

El enfoque del trabajo consistió en realizar un análisis descriptivo y predictivo de una competencia de *Kaggle* considerando una base de datos de Malasia [1], la cual tiene archivos tipo *CSV (Comma Separated Values)* de 14.993 registros de perfiles de perros y gatos con las características detalladas en la Tabla I. Esta base de datos está compuesta principalmente por variables categóricas, donde las categorías son representadas por números enteros.

Ahora bien, la variable objetivo o el valor a predecir se determina mediante la rapidez con la que se adopta una mascota o un grupo de mascotas y está dada por las siguientes categorías:

0 - Si la mascota fue adoptada el mismo día en que se incluyó en la lista.

1 - Si la mascota fue adoptada entre 1 y 7 días (1ª semana) después de su inclusión en la lista.

2 - Si la mascota fue adoptada entre 8 y 30 días (1er mes) después de su inclusión en la lista.

3 - Si la mascota fue adoptada entre 31 y 90 días (segundo y tercer mes) después de haber sido incluida en la lista.

4 - No se adoptaron las mascotas después de 100 días de estar en la lista.

Es importante resaltar que no hay mascotas en este conjunto de datos que hayan esperado entre 90 y 100 días.

De acuerdo con la información anterior, vale la pena mencionar que la variable descripción (*description*), hace referencia al texto que define a la mascota o al conjunto de mascotas y son datos utilizados para el análisis de sentimientos, particularmente el análisis de texto para identificar, extraer y estudiar información subjetiva.

Tabla I. CARACTERÍSTICAS DE LAS MASCOTAS

| Número de característica | Características de las mascotas | Tipos de datos |
|---|---|---|
| 1 | ***Type*** **=** Tipo de animal (1 = Perro, 2 = Gato) | Entero - Categórica |
| 2 | ***Name*** **=** nombre de la mascota | Cadena de texto |
| 3 | ***Age*** **=** edad de la mascota en meses | Entero |
| 4 | ***Breed1*** **=** Raza principal de mascota | Entero - Categórica |
| 5 | ***Breed2*** **=** raza secundaria de mascota, si la mascota es de raza mixta | Entero - Categórica |
| 6 | ***Gender*** **=** género de la mascota (1 = masculino, 2 = femenino, 3 = mixto, si el perfil representa un grupo de mascotas) | Entero - Categórica |
| 7 | ***Color1*** **=** color 1 de la mascota | Entero - Categórica |
| 8 | ***Color2*** **=** color 2 de la mascota | Entero - Categórica |
| 9 | ***Color3*** **=** color 3 de la mascota | Entero - Categórica |
| 10 | ***MaturitySize*** **=** tamaño (1 = pequeño, 2 = mediano, 3 = grande, 4 = extragrande, 0 = no especificado) | Entero - Categórica |
| 11 | ***FurLength*** **=** pelaje de la mascota (1 = corto, 2 = mediano, 3 = largo, 0 = no especificado) | Entero - Categórica |
| 12 | ***Vaccinated*** **=** la mascota ha sido vacunada (1 = Sí, 2 = No, 3 = No estoy seguro) | Entero - Categórica |
| 13 | ***Dewormed*** **=** la mascota ha sido desparasitada (1 = Sí, 2 = No, 3 = No estoy seguro) | Entero - Categórica |
| 14 | ***Sterilized*** **=** la mascota ha sido esterilizada / castrada (1 = Sí, 2 = No, 3 = No estoy seguro) | Entero - Categórica |
| 15 | ***Health*** **=** Condición de salud (1 = Saludable, 2 = Lesión menor, 3 = Lesión grave, 0 = No especificado) | Entero - Categórica |
| 16 | ***Quantity*** **=** número de mascotas representadas en el perfil | Entero |
| 17 | ***Fee*** **=** Tarifa de adopción (0 = Gratis) | Entero |
| 18 | ***State*** **=** ubicación del estado en Malasia | Entero - Categórica |
| 19 | ***RescuerID*** **=** ID único del rescatador | string |
| 20 | ***VideoAmt*** **=** Total de videos cargados para esta mascota | Entero |
| 21 | ***Description*** **=** descripción del perfil de cada mascota o grupo de mascotas | Cadena de texto |
| 22 | ***PetID*** **=** ID único del perfil de la mascota | Cadena de texto |
| 23 | ***PhotoAmt*** **=** Total de fotos cargadas para cada perfil de la mascota | Punto flotante |
| 24 | ***AdoptionSpeed*** **=** velocidad de adopción. Variable objetivo o valor a predecir | Entero - Categórica |

Fuente: elaboración propia (2021)

Adicionalmente, se dispone de 58.311 imágenes donde cada mascota tiene al menos una foto y cada clase tiene una cantidad determinada de imágenes por mascota, distribuidas de la siguiente manera:

- La clase cero tiene 1.363 imágenes.
- La clase uno tiene 11.517 imágenes.
- La clase dos tiene 16.438 imágenes.
- La clase tres tiene 15.059 imágenes.
- La clase cuatro tiene 13.934 imágenes.

Para iniciar se realizó la exploración de los datos con cada una de las características de las mascotas, se hizo un análisis descriptivo de cada variable con respecto a la velocidad de adopción para identificar posibles patrones de comportamiento sobre la adoptabilidad de las mascotas y con esto determinar las características que tienen mayor relación o relevancia en el valor a predecir. Luego, se realizó un preprocesamiento a las variables categóricas con una codificación de *one hot encoding*, la cual crea nuevas columnas de tipo binario por clase. Con estas variables se procede a realizar una primera ejecución de modelos de predicción.

Posteriormente, se hizo una reducción de dimensionalidad del conjunto de datos a través del análisis de componentes principales (*principal component analysis - PCA*) [13] que tiene como objetivo detectar la correlación entre variables, es decir, si existe una fuerte correlación entre las variables, el intento de reducir la dimensionalidad tiene sentido. Además, el *PCA* consiste en encontrar las direcciones de máxima varianza en datos de alta dimensión y proyectarlas a un subespacio de menor dimensión mientras se retiene la mayor parte de la información. Para complementar el análisis con las variables reducidas se pasó a entrenar y hacer la respectiva predicción con diferentes modelos como *logisticRegression*, *LGBMClassifier* y *XGBClassifier*.

En este mismo orden de ideas, se llevó a cabo la selección de características secuenciales (*Sequential Feature Selector - SFA*) [14] para tratar de reducir el error de la generalización del modelo eliminando las características irrelevantes o el ruido, y seleccionando las que tienen un mayor peso para el entrenamiento de los modelos con

*lazyClassifier* [15], la cual entrena y predice con la mayoría de modelos de aprendizaje automático para problemas de clasificación con unas métricas de *accuracy* y *f1 score* para la medición del modelo.

De igual manera, se realizó un preprocesamiento de las imágenes y se intentó ejecutar la totalidad de las mismas, pero dados los limitantes de la capacidad de cómputo, sólo se pudo entrenar con el 36.64% que corresponde a una cifra de 21.363 imágenes, realizando un modelo con 2 capas de redes neuronales convolucionales con funciones de activación relu, 32 y 64 filtros y tamaño del *kernel* de (3,3) y (2,2), además se incluyeron 2 capas de *maxpooling2D* para reducir la muestra de la entrada a lo largo de las dimensiones espaciales de alto y ancho, una capa de *flatten* para aplanar el tensor, una capa de *dropout* del 0.5 para evitar el sobreajuste del modelo y finalmente una capa densa con el número de clases con una función de activación *softmax* para la salida multiclase; luego se compiló la creación del modelo con una métrica de precisión, un optimizador *adam* y una pérdida de *categorical crossentropy*. Posteriormente, se entrenó el modelo con 25 épocas, tomando en paralelo 500 muestras. Por último, se evalúa el modelo y se sacan las predicciones para las 5 clases definidas para la velocidad de adopción de mascotas.

Finalmente, se realizó un análisis de sentimientos con el texto que se tenía asociado a la descripción de las mascotas, al cual se le aplicó un preprocesamiento con una tokenización para la variable de entrenamiento y prueba, la cual consistió en separar las palabras y darle un valor o índice a cada palabra para convertir estas oraciones en un arreglo de números. Luego, se realizó un *one hot encoding* a las variables de salida, tanto para el entrenamiento como para la prueba. Posteriormente, se creó el modelo con una capa de *embedding* que inicia la representación de los textos en números con un *max features* de 392.175, un *embedding size* de 50 y una longitud de entrada de 10; asimismo se aplicó una regularización, después se puso una capa de *Long short-term memory (LSTM)* que es una red neuronal recurrente y finalmente se añadió una capa densa con una función de activación *softmax*. Es así como se desarrolló el entrenamiento y la evaluación del modelo donde se obtuvo la predicción para las cinco clases.

# VIII. RESULTADOS

En este apartado se mostrarán los resultados obtenidos del análisis de las variables que componen el *dataset* de datos estructurados y los resultados de los modelos de aprendizaje automático y aprendizaje profundo.

## A. *ANÁLISIS DESCRIPTIVO*

Dentro del análisis descriptivo primero se observa el comportamiento de la variable objetivo velocidad de adopción (*AdoptionSpeed*), en este análisis se evidencia que algunas mascotas fueron adoptadas de inmediato. Esto corresponde a la menor proporción de los datos, con una participación del 2.73%, esto puede ser resultado del hecho de que alguien quería adoptar cualquier mascota, o la mascota tuvo la suerte de ser vista por una persona que quería dicho perfil. La generalidad es que las mascotas son adoptadas posterior a los 8 días de estar incluidas en la plataforma, esto se puede observar en la Fig 1.

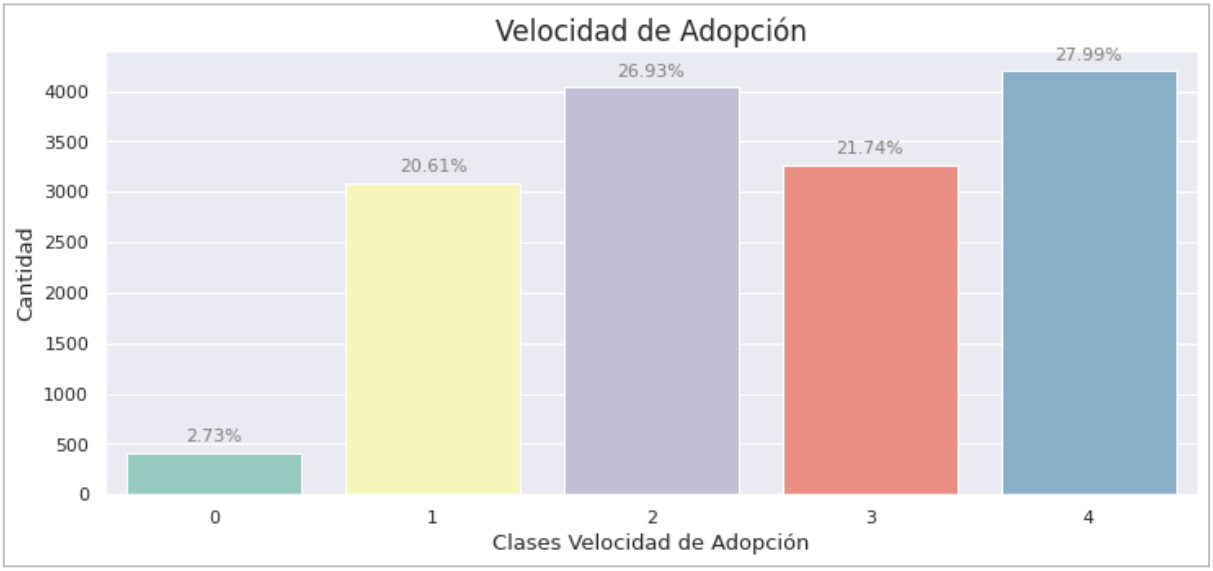


Fig 1. Velocidad de adopción de las mascotas

Fuente: kaggle. (2018) y elaboración propia (2021)

Ahora bien, para el tipo de mascota se muestra que el mayor porcentaje de los perros se adoptan después de los 100 días, mientras que los gatos se adoptan entre 8 y 30 días. Esto puede indicar que los gatos son adoptados en menor tiempo, sin embargo, las categorías de velocidad de adopción se muestran muy similares para ambas mascotas, indicando que hay

una mayor cantidad de perros que de gatos en los datos. Esto se puede visualizar en la Fig. 2.

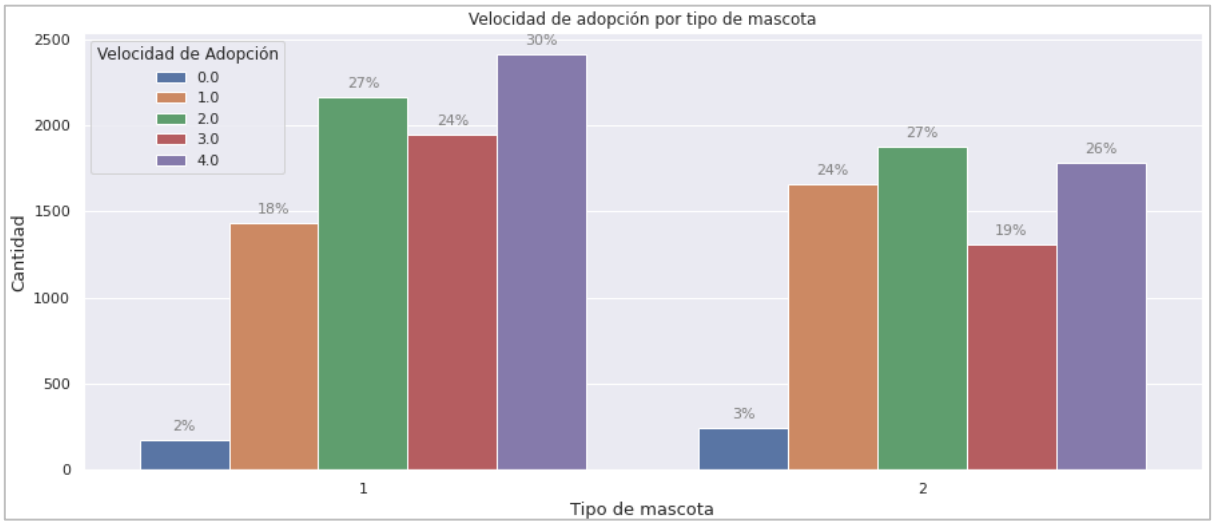


Fig 2. Velocidad de adopción por tipo de mascota

Fuente: kaggle. (2018) y elaboración propia (2021)

Para la variable "nombre" se tiene que hay pocas mascotas que no tienen nombre dentro del conjunto de datos y las que tienen un nombre definido se adoptan en mayor proporción el primer mes o a los 100 días después de estar incluidas en la lista de adopción, esto se evidencia en la Fig. 3.

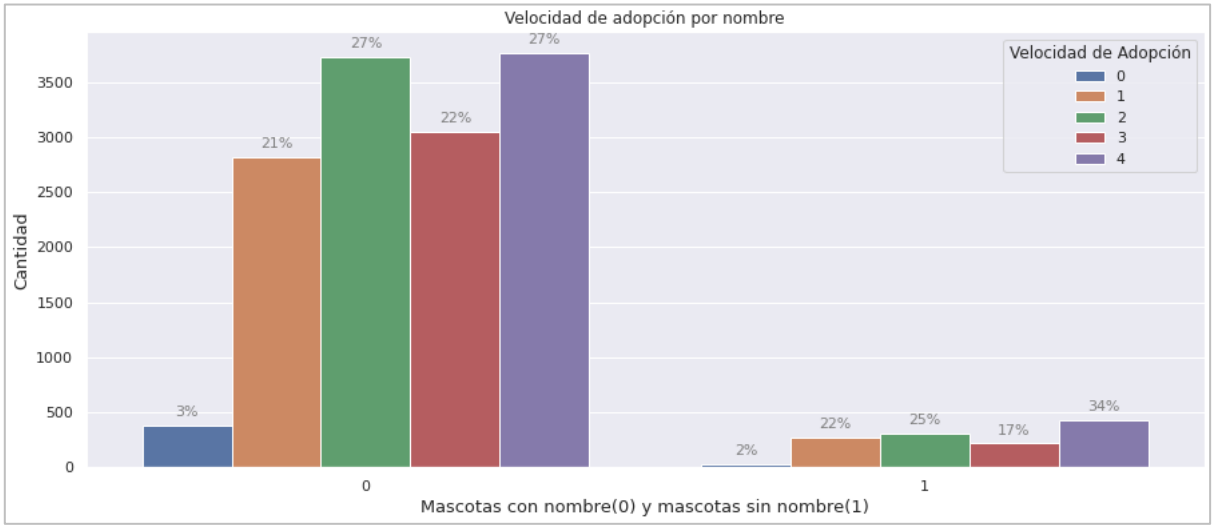


Fig 3. Velocidad de adopción por nombre

Fuente: kaggle. (2018) y elaboración propia (2021)

Para el atributo de las razas principales se obtiene el top de las 5, las cuales son:

1. Raza mixta (*Mixed Breed*)
2. Pelo corto (*Domestic Short Hair)*
3. Pelo medio *(Domestic Medium Hair)*
4. Gato atigrado *(Tabby)*
5. Pelo largo *(Domestic Long Hair)*

También se obtiene el top 5 de las razas secundarias:

1. No tiene identificada una raza
2. Raza mixta *(Mixed Breed)*
3. Pelo corto (*Domestic Short Hair)*
4. Pelo medio *(Domestic Medium Hair)*
5. Gato atigrado *(Tabby)*

De acuerdo con la información encontrada se puede concluir que la mayoría de mascotas para adoptar tienen razas mezcladas y no tienen una raza identificada.

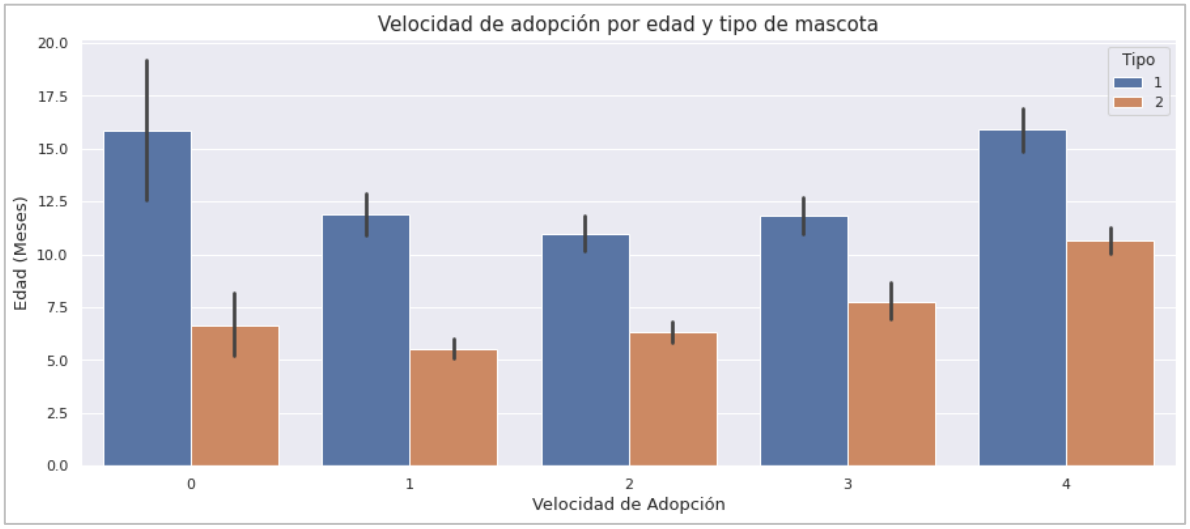


Fig 4.Velocidad de adopción por edad y tipo de mascota

Fuente: kaggle. (2018) y elaboración propia (2021)

También es importante tener en cuenta la edad por tipo de mascotas, en esta se observa una alta variabilidad de la edad debido a la extensión de la línea a partir de la mediana, esto se visualiza en la Fig. 4.

Además, se puede observar que la edad de los perros tiene mayor variabilidad con respecto a los gatos, entre menos edad tengan los animales del perfil se evidencia mayor velocidad de adopción.

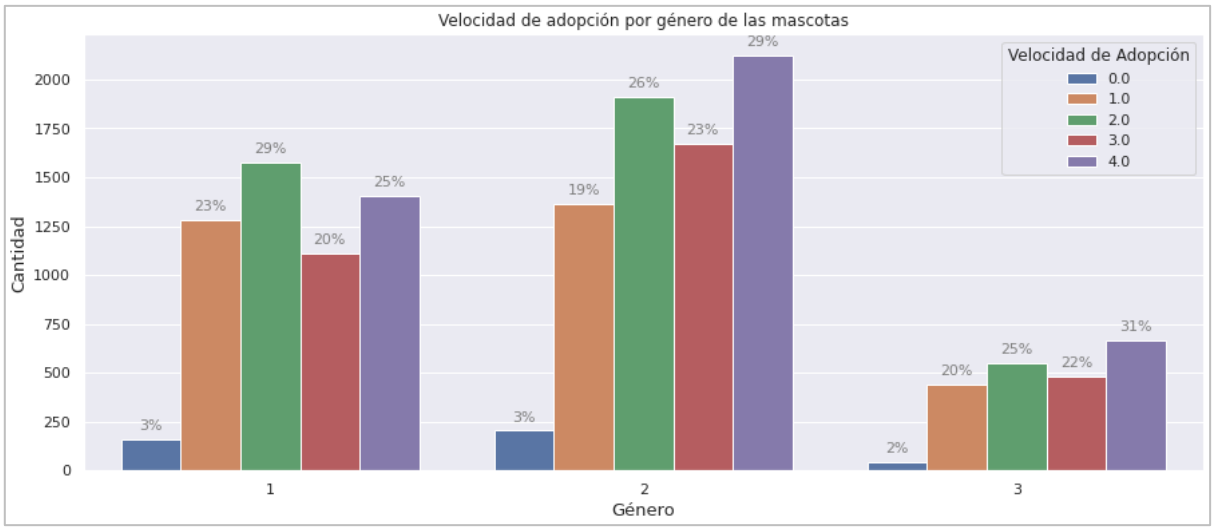


Fig 5.Velocidad de adopción por género de las mascotas

Fuente: kaggle. (2018) y elaboración propia (2021)

En la Fig. 5 se logra evidenciar que hay una mayor proporción de hembras y que estas se adoptan después de los 100 días de estar en la lista de adopción, aunque no hay mucha diferencia entre las clases 1,2 y 3; por lo que el sexo no tiene una incidencia clara en la velocidad de adopción, puesto que hay similitud en el comportamiento entre las categorías analizadas.

En la Fig. 6, se identifica que la proporción de mascotas de pelaje corto es mayor con respecto a las demás categorías y la mayor cantidad se adoptan después de los 100 días de estar incluidas en la lista.

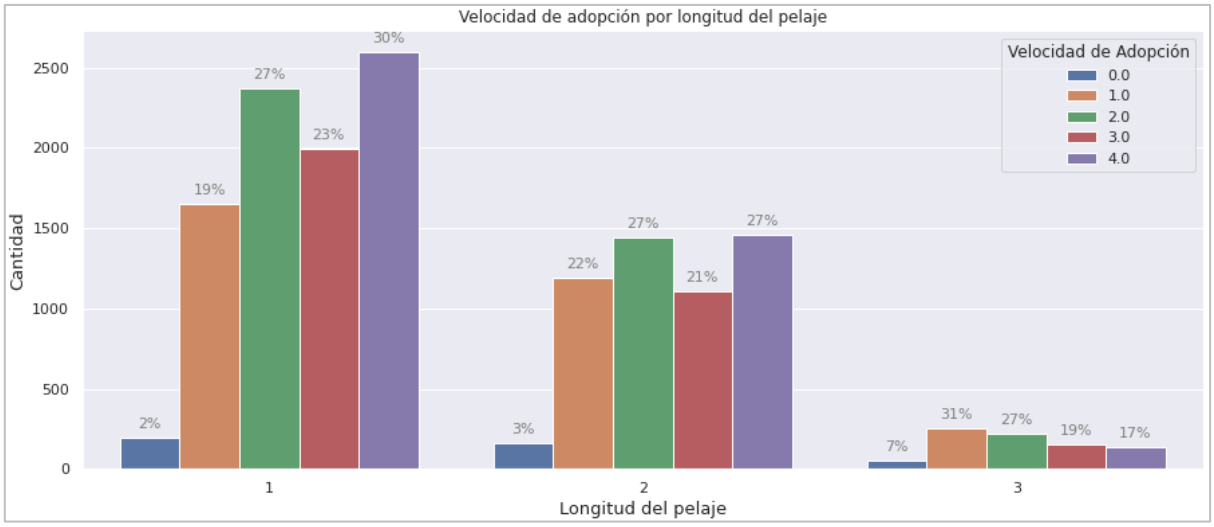


Fig 6.Velocidad de adopción por longitud del pelaje de las mascotas

Fuente: kaggle. (2018) y elaboración propia (2021)

En la Fig. 7 se observa que el tipo de animal mediano tiene mayor participación en la base de datos donde la mayor cantidad de éstos son adoptados después de los 100 días de estar incluidos en la lista. Cabe anotar que el tipo de animales extragrande no tiene participación relevante en la base de datos.

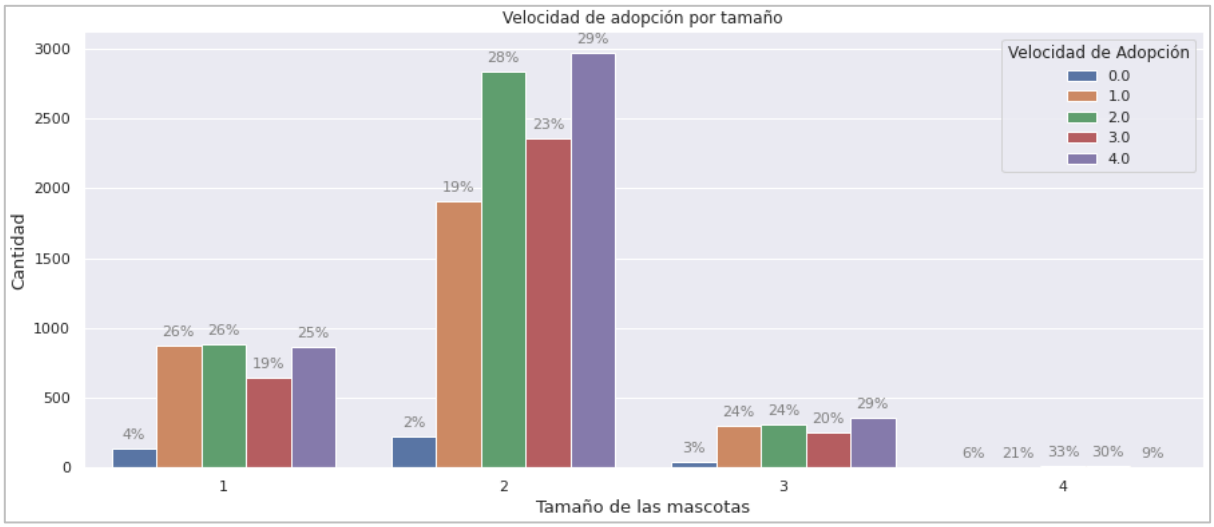


Fig 7.Velocidad de adopción por tamaño de las mascotas

Fuente: kaggle. (2018) y elaboración propia (2021)

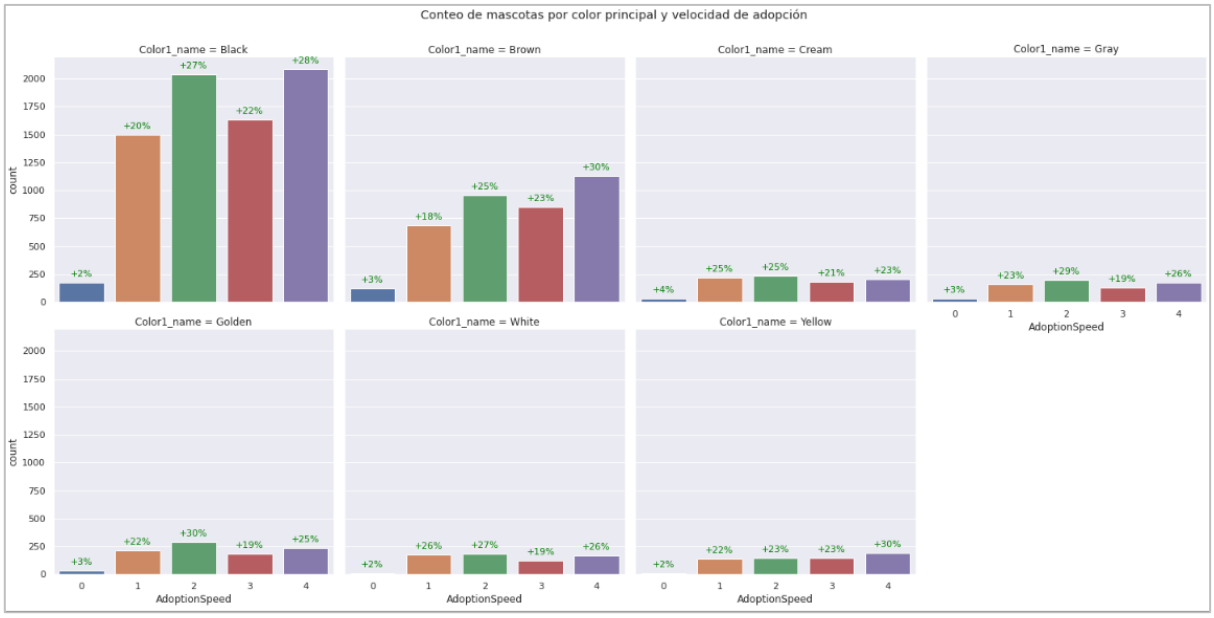


Fig 8.Conteo de mascotas por color principal y velocidad de adopción

Fuente: kaggle. (2018) y elaboración propia (2021)

Cuando se analiza el color con respecto a la velocidad de adopción en la Fig. 8, se evidencia que el color predominante en las mascotas es el negro, y que estas mascotas se adoptan más en la categoría 2 y 4, es decir, entre 8 y 30 días y después de 100 días de estar en la lista de adopción.

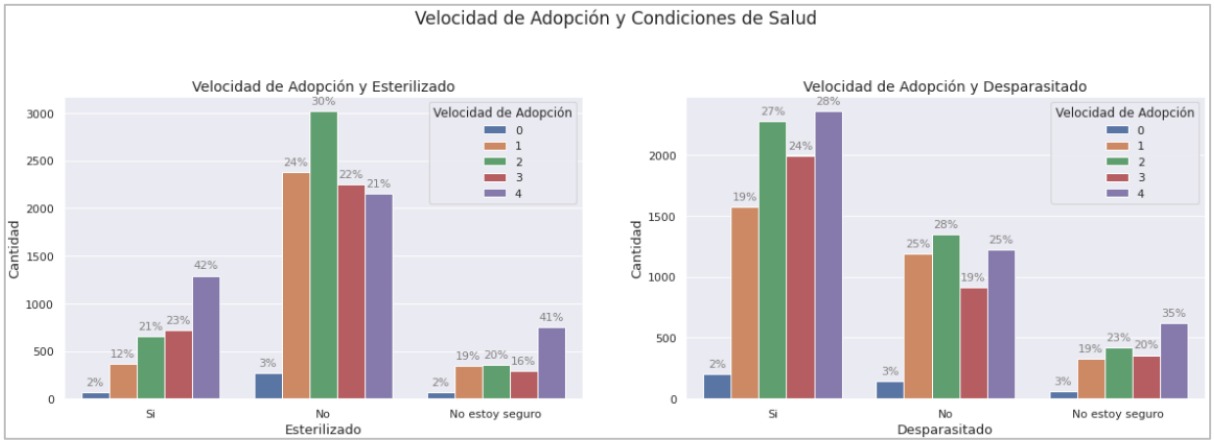


Fig 9.Velocidad de adopción y condiciones de salud de las mascotas

Fuente: kaggle. (2018) y elaboración propia (2021)

De acuerdo con la Fig. 9 se puede inferir que las personas prefieren mascotas sin esterilizar, quizás porque desean tener cachorros o gatos pequeños. Además, la mayoría de estas mascotas se adoptan en el primer mes de incluirlos en la lista de adopción. También, es

importante mencionar que se adoptan fácilmente las mascotas desparasitadas que llevan en lista entre 30 y 100 días.

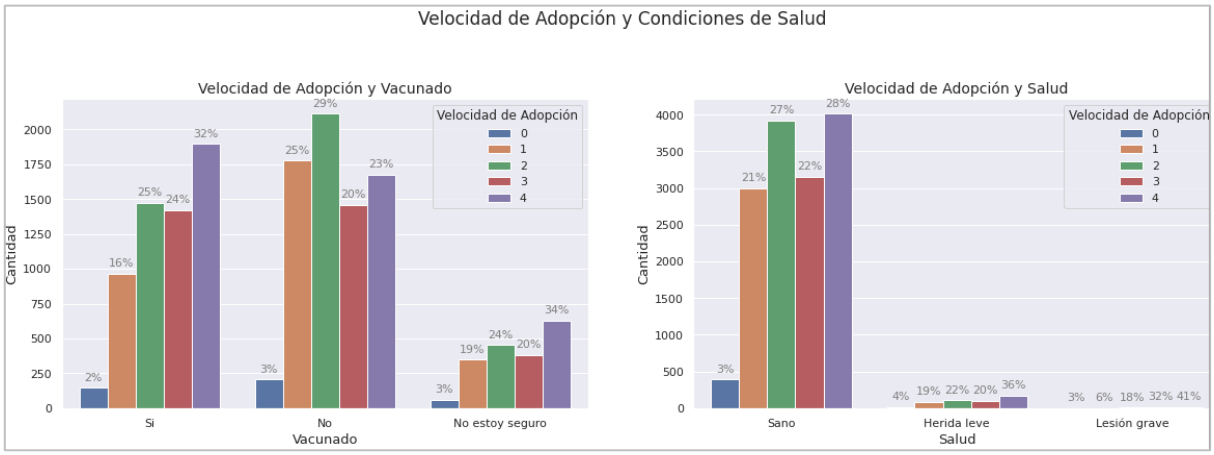


Fig 10.Velocidad de adopción y condiciones de salud de las mascotas

Fuente: kaggle. (2018) y elaboración propia (2021)

Para la Fig. 10 se muestra que las personas prefieren adoptar mascotas no vacunadas. Estas mascotas se adoptan después de estar un mes en la lista de adopción. Igualmente, se dan en adopción los animales que están sanos y que llevan entre un mes y menos de 90 días en la lista. Aunque la mayoría de las mascotas están sanas.

Finalmente, es importante mencionar que cuando no se tiene información sobre el estado de salud de la mascota, la probabilidad de no ser adoptado es mucho más alta, y cuando se tienen mascotas sanas, desparasitadas y no esterilizadas tienden a ser adoptadas más rápido.

Dentro de este mismo orden de ideas, se analizó la cantidad de mascotas por perfil seleccionando los grupos de animales mayores a 11 y con esto se logró llegar a lo siguiente, en la Fig. 11 se muestra que la mayor cantidad de mascotas se adoptan entre los 8 y 30 días, y después de los 100 días de ser incluidas en la lista de adopción. Además, se puede observar que el menor número de mascotas se adoptan la primera semana.

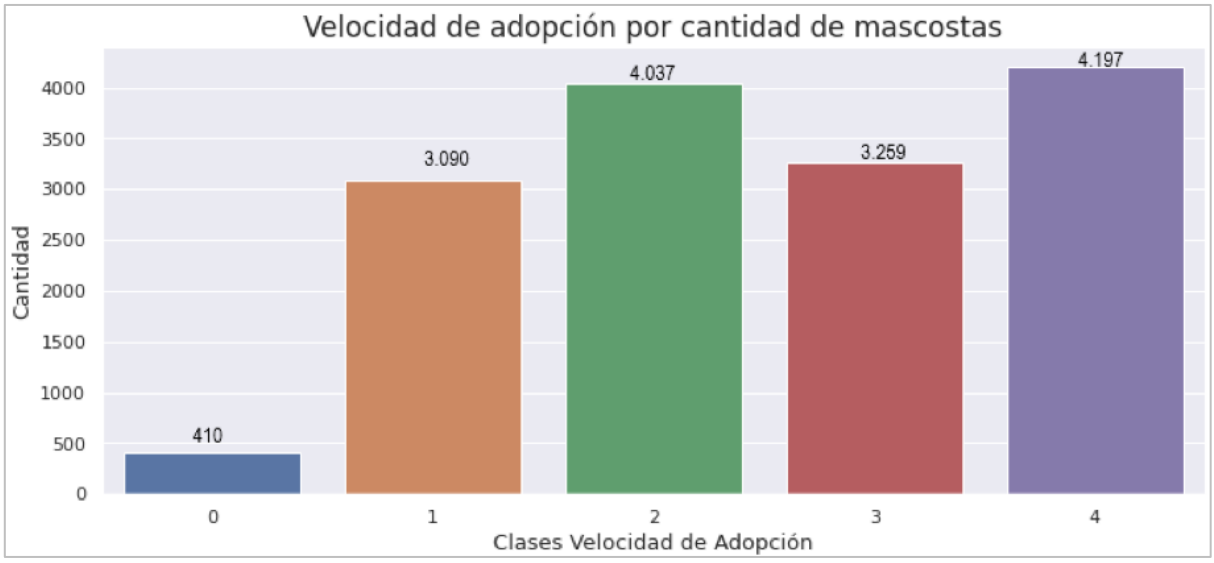


Fig 11.Cantidad de mascotas representadas en el perfil

Fuente: kaggle. (2018) y elaboración propia (2021)

Para el atributo de tarifa es importante aclarar que algunas mascotas se pueden conseguir de forma gratuita, aunque para la adopción de otras es necesario pagar una cierta cantidad de dinero. Sin embargo, la mayoría de las mascotas se entregan gratis y las tarifas suelen ser inferiores a $100. Esto se observa en la Fig. 12.

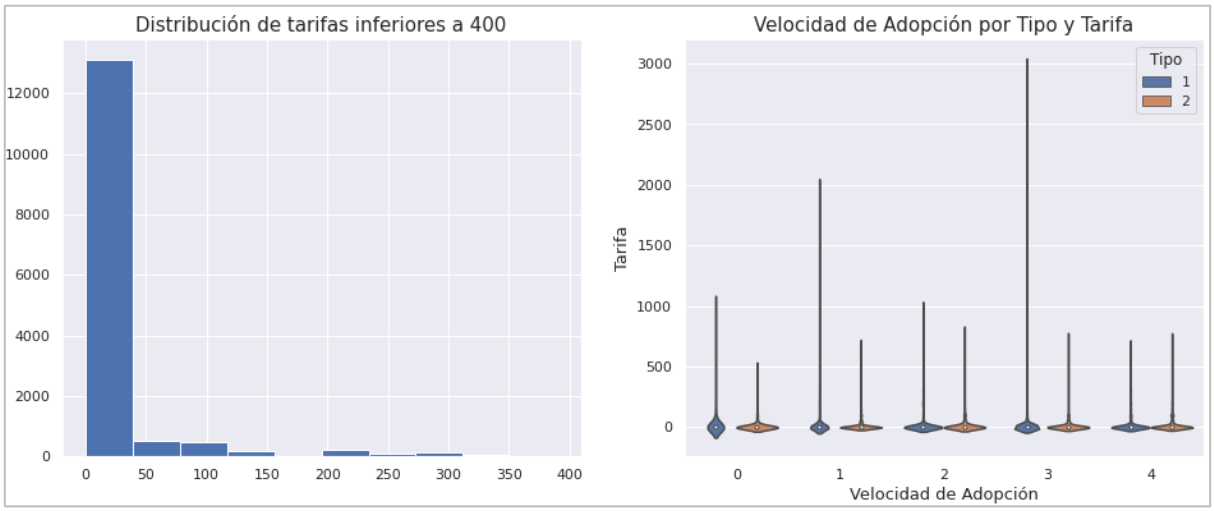


Fig 12. Velocidad de adopción por tarifa y tipo de mascota

Fuente: kaggle. (2018) y elaboración propia (2021)

Las mascotas con tarifas altas tienden a ser adoptadas bastante rápido, tal vez la gente prefiera pagar por mascotas "mejores", sanas, entrenadas, entre otros aspectos. Además, las tarifas para los perros tienden a ser más altas, aunque estos son casos atípicos.

Para complementar las gráficas anteriores se hace el siguiente análisis que tiene en cuenta la cantidad de mascotas por tipo y tarifa.

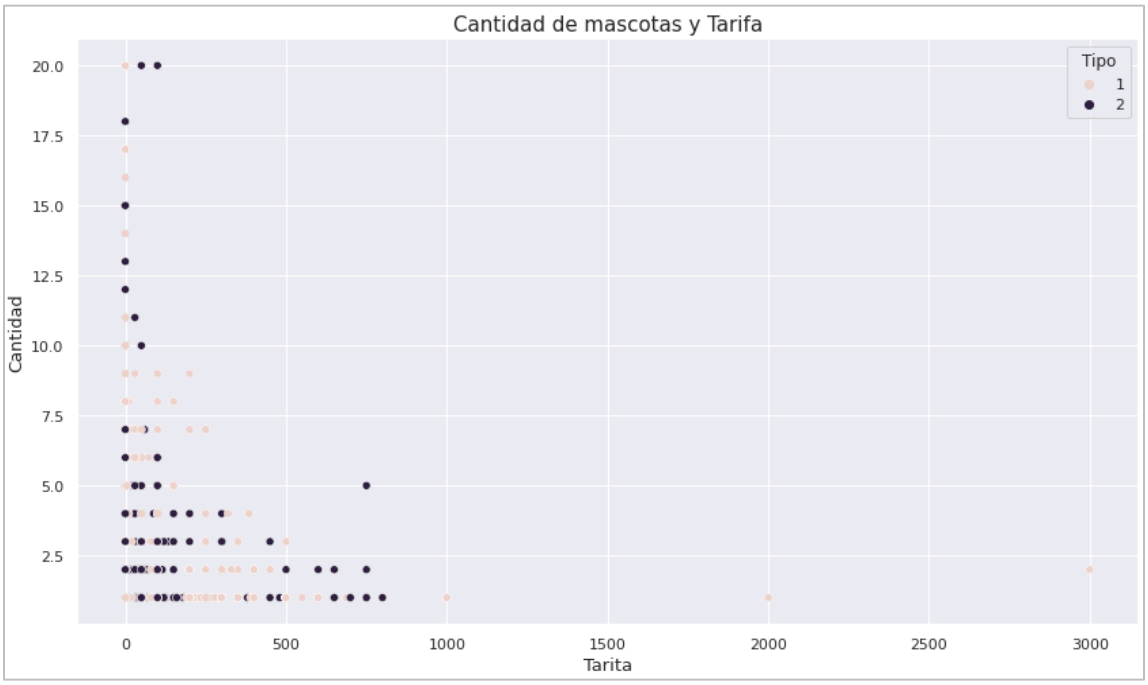


Fig 13. Cantidad de mascotas por tipo y tarifa

Fuente: kaggle. (2018) y elaboración propia (2021)

Según la Fig.13, las tarifas y la cantidad de mascotas tienen una relación inversamente proporcional, ya que entre menos mascotas haya, mayor es la tarifa, quizá porque estas mascotas solas están mejor entrenadas y preparadas que las demás. Cabe resaltar que las mascotas que se venden a un precio más alto son los perros y hay menor cantidad de estos casos.


Fig 14. Cantidad de mascotas por distribución geográfica y velocidad de adopción

Fuente: kaggle. (2018) y elaboración propia (2021)

Según la Fig. 14 y la Tabla II, el estado de *Selangor* de Malasia es donde hay mayor velocidad de adopción debido a la cantidad de anuncios.

Tabla II. ANUNCIOS POR ESTADOS DE MALASIA

| **Estado de Malasia** | **Cantidad de anuncios de adopción de mascotas (%)** |
|---|---|
| Selangor | 57% |
| Kuala Lumpur | 26% |
| Pulau Pinang | 7,13% |
| Johor | 3,34% |
| Perak | 2,99% |

Fuente: kaggle. (2018) y elaboración propia (2021)

De acuerdo con la relación entre los atributos de las mascotas y la velocidad de adopción, la mayoría de características tienen una relación inversamente proporcional con respecto a la variable objetivo. No obstante, los atributos que tienen un mayor peso son la edad (*age)*, la raza principal *(Breed1)* y la esterilización *(Sterilized)*, presentan una mayor relación con la velocidad de adopción, esto se evidencia en la Tabla III.

Tabla III. VALORES CORRELACIÓN VARIABLES VS VELOCIDAD DE ADOPCIÓN

| | **Valor correlación con *AdoptionSpeed*** |
|---|---|
| *Type* | -0,091 |
| *Age* | 0,101 |
| *Breed1* | 0,108 |
| *Breed2* | -0,019 |
| *Gender* | 0,058 |
| *Color1* | -0,044 |
| *Color2* | -0,039 |
| *Color3* | -0,007 |
| *MaturitySize* | 0,046 |
| *FurLength* | -0,091 |
| *Vaccinated* | -0,059 |
| *Dewormed* | -0,013 |
| *Sterilized* | -0,083 |
| *Health* | 0,029 |
| *Quantity* | 0,063 |
| *Fee* | -0,004 |
| *State* | 0,013 |
| *VideoAmt* | -0,001 |
| *PhotoAmt* | -0,023 |

Fuente: kaggle. (2018) y elaboración propia (2021)

## *B. RESULTADO DE LOS MODELOS PREDICTIVOS*

Con la base de datos de entrenamiento de la competencia de *Kaggle (Train.csv)* que considera los datos estructurados se realizan dos iteraciones para diferentes modelos, la primera iteración considerando la totalidad de variables que se pueden entregar a los modelos de aprendizaje automático clásicos. En esta se consideran las siguientes variables: *Type, Age, Quantity, Fee, VideoAmt, PhotoAmt, Breed1, Breed2, Gender, Color1, Color2, Color3, MaturitySize, FurLength, Vaccinated, Dewormed, Sterilized, Healthy State.* No se consideran variables como *Name, RescuerID, Description* y *PetID* ya que son valores únicos por perfil y no entregan información que aporte a predecir la velocidad de adopción mediante los modelos clásicos. Se realiza un *one hot encoding* para las características que tienen asignados valores enteros por clase. Posteriormente se procede a ejecutar la librería de *LazyPredict* que

permite el ajuste y evaluación de los modelos contenidos en la librería de *scikit-learn*, éste entrega en orden los modelos que mejor valor de precisión y *F1-score* se obtienen.

El resultado de los mejores algoritmos para la primera iteración se puede visualizar en la Tabla IV, el primer modelo relacionado es el *XGBClassifier* [16] el cual corresponde a una implementación popular del aumento del gradiente, donde se entrena minimizando la pérdida de la función objetivo frente a un conjunto de datos. Este modelo comienza con una predicción inicial y se usa la función de pérdida para evaluar la predicción, a partir de esta construye un árbol que busca minimizar la pérdida iterando en las diferentes ramas del árbol hasta encontrar el valor óptimo. El segundo modelo relacionado es el *LGBMClassifier* [17] el cual divide el árbol a partir del mejor ajuste, esto ayuda a disminuir los cálculos por niveles, obteniendo una precisión mayor. El último modelo corresponde al *RandomForestClassifier* [18] que es un conjunto de árboles de decisión combinados con *bagging*, es decir, que distintos árboles están evaluando distintas porciones de datos; esto produce que cada árbol se entrene con distintas muestras de datos para el problema, con esto, al combinar los resultados los errores se compensan y se obtiene una predicción general adecuada.

Tabla IV. RESULTADOS ITERACIÓN 1

| | | | **Estadísticas** | | |
|---|---|---|---|---|---|
| | | | **Precisión** | ***F1 - score*** | **Variables consideradas** |
| **Modelos Aprendizaje Automático** | **Iteración 1** | *XGBClassifier* | 0,41 | 0,39 | *Type, Age, Quantity, Fee, VideoAmt, PhotoAmt, Breed1, Breed2, Gender, Color1, Color2, Color3, MaturitySize, FurLength, Vaccinated, Dewormed, Sterilized, Health, State.* |
| | | *LGBMClassifier* | 0,40 | 0,38 | |
| | | *RandomForestClassifier* | 0,38 | 0,37 | |

Fuente: kaggle. (2018) y elaboración propia (2021)

Para la segunda iteración, primero se procede a realizar una selección de características y una reducción de dimensionalidad con un análisis de componentes principales (*Principal Component Analysis - PCA*). Para la selección de características se realiza la ejecución del modelo de selección secuencial hacia adelante con el objetivo de reducir el espacio de d-dimensiones, considerando diecisiete características, a un espacio de k-dimensiones, donde k<d, buscando aquellas que son relevantes para el problema. Este modelo se ejecuta con el algoritmo k vecinos más cercanos o *KNN (K-Nearest Neighbours)*,

el cual a partir de los datos iniciales clasifica las instancias nuevas. En la Tabla V se relacionan los valores obtenidos para cada grupo de características, estos resultados indican que el número de variables a considerar es 12 y también entrega las que se deben considerar, las cuales corresponden a: *Sterilized*, *Breed1*, *Age*, *MaturitySize*, *Quantity*, *PhotoAmt*, *Vaccinated, State, FurLength, Health, Fee, Dewormed.*

Tabla V. RESULTADOS SELECCIÓN SECUENCIAL HACIA ADELANTE

| **Número de características** | **Precisión** | **Características** |
|---|---|---|
| 1 | 0,2775 | *Sterilized* |
| 2 | 0,2814 | *Sterilized , Breed1* |
| 3 | 0,2702 | *Sterilized , Breed1, Age* |
| 4 | 0,2981 | *Sterilized , Breed1, Age, MaturitySize* |
| 5 | 0,3167 | *Sterilized , Breed1, Age, MaturitySize, Quantity* |
| 6 | 0,3276 | *Sterilized , Breed1, Age, MaturitySize, Quantity, PhotoAmt* |
| 7 | 0,3289 | *Sterilized , Breed1, Age, MaturitySize, Quantity, PhotoAmt, Vaccinated* |
| 8 | 0,3292 | *Sterilized , Breed1, Age, MaturitySize, Quantity, PhotoAmt, Vaccinated, State* |
| 9 | 0,3399 | *Sterilized , Breed1, Age, MaturitySize, Quantity, PhotoAmt, Vaccinated, State, FurLength* |
| 10 | 0,3408 | *Sterilized , Breed1, Age, MaturitySize, Quantity, PhotoAmt, Vaccinated, State, FurLength, Health* |
| 11 | 0,3438 | *Sterilized , Breed1, Age, MaturitySize, Quantity, PhotoAmt, Vaccinated, State, FurLength, Health, Fee* |
| 12 | 0,344 | *Sterilized , Breed1, Age, MaturitySize, Quantity, PhotoAmt, Vaccinated, State, FurLength, Health, Fee, Dewormed* |
| 13 | 0,3401 | *Sterilized , Breed1, Age, MaturitySize, Quantity, PhotoAmt, Vaccinated, State, FurLength, Health, Fee, Dewormed, Color1* |
| 14 | 0,3419 | *Sterilized , Breed1, Age, MaturitySize, Quantity, PhotoAmt, Vaccinated, State, FurLength, Health, Fee, Dewormed, Color1, Type* |
| 15 | 0,3408 | *Sterilized , Breed1, Age, MaturitySize, Quantity, PhotoAmt, Vaccinated, State, FurLength, Health, Fee, Dewormed, Color1, Type, Gender* |
| 16 | 0,3332 | *Sterilized , Breed1, Age, MaturitySize, Quantity, PhotoAmt, Vaccinated, State, FurLength, Health, Fee, Dewormed, Color1, Type, Gender, Color2* |
| 17 | 0,3254 | *Sterilized , Breed1, Age, MaturitySize, Quantity, PhotoAmt, Vaccinated, State, FurLength, Health, Fee, Dewormed, Color1, Type, Gender, Color2, Color3* |

Fuente: kaggle. (2018) y elaboración propia (2021)

Para la reducción de la dimensionalidad con el análisis de componentes principales (*Principal Component Analysis - PCA*), se realiza la ejecución para tres tipos de modelos clásicos: *LogisticRegression, LGBMClassifier y XGBClassifier;* estos se evalúan a partir de la métrica de precisión, la cual indica que no se debe realizar reducción de dimensionalidad ya que las variables no se encuentran relacionadas, por tal motivo los mejores valores de precisión para los modelos considerados indican que se deben tener en cuenta las 17 variables que se tienen como entrada. Esto se puede visualizar en la Tabla VI, en esta se evidencia la iteración por el número de variables a reducir y el valor de precisión para cada número.

Tabla VI. RESULTADOS ANÁLISIS DE COMPONENTES PRINCIPALES

| **Modelo** | **Resultados de precisión por número de características final** |
|---|---|
| *LogisticRegression* | **1:** 0.297, **2:** 0.299, **3:** 0.297, **4:** 0.316, **5:** 0.318, **6:** 0.319, **7:** 0.324, **8:** 0.327, **9:** 0.330, **10:** 0.330, **11:** 0.328, **12:** 0.329, **13:** 0.327, **14:** 0.329, **15:** 0.347, **16:** 0.349, **17: 0.357** |
| *LGBMClassifier* | **1:** 0.267, **2:** 0.287, **3:** 0.314, **4:** 0.325, **5:** 0.325, **6:** 0.327, **7:** 0.331, **8:** 0.342, **9:** 0.352, **10:** 0.353, **11:** 0.358, **12:** 0.359, **13:** 0.361, **14:** 0.358, **15:** 0.376, **16:** 0.376, **17: 0.380** |
| *XGBClassifier* | **1:** 0.283, **2:** 0.300, **3:** 0.319, **4:** 0.330, **5:** 0.322, **6:** 0.324, **7:** 0.334, **8:** 0.333, **9:** 0.342, **10:** 0.355, **11:** 0.349, **12:** 0.360, **13:** 0.360, **14:** 0.361, **15:** 0.368, **16: 0.374**, **17:** 0.371 |

Fuente: kaggle. (2018) y elaboración propia (2021)

Finalmente, con el objetivo de mejorar la primera iteración se procede a ejecutar el comando de *LazyPredict* con las 12 variables que arrojó como resultado la selección de características; en estos resultados se evidencia que los valores se mantienen constantes, lo que indica que se tenían 7 variables que no le estaban aportando a los modelos. Los resultados de la segunda iteración se pueden observar en la Tabla VII.

Tabla VII. RESULTADOS ITERACIÓN 2

| | | | **Estadísticas** | | |
|---|---|---|---|---|---|
| | | | **Precisión** | ***F1 - score*** | **Variables consideradas** |
| **Modelos Aprendizaje Automático** | **Iteración 2** | *XGBClassifier* | 0,41 | 0,39 | *Sterilized , Breed1, Age, MaturitySize, Quantity, PhotoAmt, Vaccinated, State, FurLength, Health, Fee, Dewormed* **** Sequential Feature Selector*** |
| | | *LGBMClassifier* | 0,41 | 0,39 | |
| | | *RandomForestClassifier* | 0,38 | 0,38 | |

Fuente: kaggle. (2018) y elaboración propia (2021)

La tercera iteración del problema corresponde a la ejecución de los modelos de aprendizaje profundo. Primero se realiza el preprocesamiento para 21.363 imágenes, esto es una muestra del total de imágenes ya que por capacidad de computo no fue posible ejecutar el modelo para el *dataset* completo. Se construye una arquitectura de clasificación de imágenes con 2 capas de redes neuronales convolucionales con funciones de activación *relu*, 32 y 64 filtros y tamaño del *kernel* de (3,3) y (2,2); 2 capas de *maxpooling2D*; una capa de *flatten*; una capa de *dropout* del 0.5 y finalmente una capa densa con el número de clases con una función de activación *softmax* para la salida multiclase; se compila el modelo con una métrica de precisión, un optimizador *adam* y una pérdida de *categorical crossentropy*. Para el análisis de sentimientos con el texto que se tiene asociado a la descripción de las mascotas, se aplicó un preprocesamiento con una tokenización, luego, se realizó un *one hot encoding* a las variables de salida. Posteriormente, se creó el modelo con una capa de *embedding*, un *embedding size* de 50 y una longitud de entrada de 10; asimismo se aplicó una regularización, se montó una capa de *Long short-term memory (LSTM)* y finalmente se añadió una capa densa con una función de activación *softmax*. Los resultados obtenidos a partir de estas arquitecturas se pueden percibir en la Tabla VIII. La precisión obtenida para estos modelos es baja debido a que el problema de clasificación planteado es complejo, ya que la clasificación que deben realizar estos algoritmos es difícil incluso para el cerebro humano, ya que los datos no estructurados asociados a cada perfil no tienen diferencias de fácil percepción entre las clases.

Los modelos de aprendizaje profundo pueden ser mejorados con variaciones de las capas de entrenamiento utilizadas o mediante técnicas como el *transfer learning*, esta consiste en tomar características aprendidas de un modelo y aprovecharlas en un problema similar, el flujo de trabajo de este tipo de técnicas consiste en tomar capas de un modelo previamente entrenado, inmovilizar estas para evitar destruir la información en las rondas de entrenamiento posteriores, agregar las capas nuevas y entrenables sobre las capas inmovilizadas y finalmente se entrenar las nuevas capas en el conjunto de datos seleccionado. El último paso el cual es opcional es el *fine-tuning* que consiste en realizar el entrenamiento con todas las capas incluso las inmovilizadas inicialmente, esto con los nuevos datos y una tasa de aprendizaje muy baja [19].

Tabla VIII. RESULTADOS ITERACIÓN 3

| | | | **Precisión** |
|---|---|---|---|
| **Modelos Aprendizaje Profundo** | **Iteración 3** | Clasificación de imágenes | 0,24 |
| | | Clasificación de texto | 0,28 |

Fuente: kaggle. (2018) y elaboración propia (2021)

Pese a la complejidad del problema abordado, los resultados obtenidos superan el estado del arte construido por Zhang et al. [12], este trabajo logra encontrar modelos que realizan mejores predicciones mediante la óptica de los modelos de aprendizaje automático clásicos, obteniendo modelos con precisión del 41%, superando el 38% que indican estos autores.

## IX. CONCLUSIONES

En la predicción de la velocidad de adopción de las mascotas influyen las características de raza principal, cantidad de mascotas por perfil, esterilización, tamaño de la mascota, vacunación, estado donde se encuentra ubicada la mascota, estado de salud, tarifa, pelaje, desparasitación, edad y la cantidad de fotos asociadas en el perfil. Estas variables se obtienen a partir de la selección de características secuencial hacia adelante con la implementación de un modelo de k vecinos más cercanos con una precisión del 34,40% sobre la velocidad de adopción, permitiendo eliminar 7 características que no aportan a los modelos de aprendizaje automático clásicos.

Dada las técnicas utilizadas de modelos tradicionales de *machine learning* como la *LogisticRegression*, *LGBMClassifier*, *XGBClassifier* y el *RandomForestClassifier* independientemente de la cantidad de variables que se tengan en cuenta, el modelo que arrojó una mayor precisión fue el *XGBClassifier* con un 41% para una primera iteración. Luego, para la segunda iteración este modelo se igualó en resultados con el *LGBMClassifier*, dando como resultado una precisión del 41%.

Para la implementación de los modelos de *deep learning*, en la tercera iteración con el análisis de las imágenes se obtuvo una precisión del 24% y para la cuarta iteración que fue el análisis de los textos se consiguió un resultado del 28%, corroborando que se trata de un problema complejo, ya que la clasificación que deben realizar estos algoritmos es difícil incluso para el cerebro humano.

Dados los resultados anteriores, las características de las mascotas parecen tener una mayor influencia en la decisión de adopción. No obstante, los resultados obtenidos indican que el problema es complejo, ya que la exactitud de las predicciones tiene valores bajos en las diferentes iteraciones del modelo, debido a que es muy difícil reconocer qué atributos o características diferencian a una mascota de otra para ser adoptado en determinado tiempo, dado que no hay un patrón, los modelos son imprecisos y no logran realizar una mejor predicción.

De igual manera, es importante mencionar que los modelos de *machine learning* superan el estado del arte existente para el problema abordado, Zhang et al. [12] obtuvieron una

precisión para el *Random Forest* del 38%; aunque no se lograron superar los resultados de los modelos de *deep learning* que fueron del 39%, lo cual genera nuevos retos y alternativas para mejoras en los modelos predictivos de *deep learning.*

## X. REFERENCIAS


[1] "PetFinder.my Adoption Prediction," 2018. https://www.kaggle.com/c/petfinder-adoption-prediction/data?select=StateLabels.csv (accessed Nov. 21, 2021).

[2] "Malasia: Economía y demografía," 2021. https://datosmacro.expansion.com/paises/malasia (accessed Nov. 21, 2021).

[3] CESCE - El valor del crédito, "INFORME RIESGO PAÍS," 2021. Accessed: Nov. 21, 2021. [Online]. Available: https://www.cesce.es/documents/20122/0/INFORME++MALASIA+-+26+mayo+2021.pdf/c5e78026-b39b-0e36-f246-d6e5732d99e8?t=1622539754625

[4] "Animal Welfare Act 2015 Gazetted," Jan. 21, 2016. https://www.spca.org.my/news/animal-welfare-act-2015-gazetted/ (accessed Nov. 21, 2021).

[5] NOORSILA ABD MAJID, "Malaysia announces new animal welfare law with more bite," Jul. 18, 2017. https://www.nst.com.my/news/nation/2017/07/258472/malaysia-announces-new-animal-welfare-law-more-bite#:~:text=PUTRAJAYA%3A%20Animal%20abusers%20beware.&text=%22Under%20this%20Act%2C%20those%20who,action%2C%22%20warned%20Ahmad%20Shabery. (accessed Nov. 21, 2021).

[6] Noel Wong, "Group boosts dog-adoption rates through social media," Sep. 17, 2019. https://www.freemalaysiatoday.com/category/leisure/2019/09/17/group-boosts-dog-adoption-rates-through-social-media/?__cf_chl_jschl_tk__=.nYdUVLBQkEUJFoCUdtII6IyPIP7UzIUThkz1xKf1r0-1637626552-0-gaNycGzNDqU (accessed Nov. 21, 2021).

[7] I. H. Sarker, "Machine Learning: Algorithms, Real-World Applications and Research Directions," *SN Computer Science*, vol. 2, no. 3, May 2021, doi: 10.1007/s42979-021-00592-x.

[8] I. Corso and C. Lorena, "Aplicación de algoritmos de clasificación supervisada usando Weka." Accessed: Nov. 21, 2021. [Online]. Available: https://www.investigacion.frc.utn.edu.ar/labsis/Publicaciones/congresos_labsis/cynthia/CNIT_2009_Aplicacion_Algoritmos_Weka.pdf

[9] S. Ghosh, N. Das, I. Das, and U. Maulik, "Understanding deep learning techniques for image segmentation," *ACM Computing Surveys*, vol. 52, no. 4, Aug. 2019, doi: 10.1145/3329784.

[10] J. Bradley and S. Rajendran, "Increasing adoption rates at animal shelters: a two-phase approach to predict length of stay and optimal shelter allocation," *BMC Veterinary Research*, vol. 17, no. 1, Dec. 2021, doi: 10.1186/s12917-020-02728-2.

[11] J. Alcaidinho *et al.*, "Leveraging mobile technology to increase the permanent adoption of shelter dogs," in *MobileHCI 2015 - Proceedings of the 17th International Conference on*

*Human-Computer Interaction with Mobile Devices and Services*, Aug. 2015, pp. 463–469. doi: 10.1145/2785830.2785861.

[12] K. Zhang and S. Zhang, "PetFinder Challenge: Predicting Pet Adoption Speed." Accessed: Nov. 21, 2021. [Online]. Available: http://cs229.stanford.edu/proj2019spr/report/55.pdf

[13] "Principal Component Analysis," 2020. http://rasbt.github.io/mlxtend/user_guide/feature_extraction/PrincipalComponentAnalysis/ (accessed Nov. 21, 2021).

[14] "Sequential Feature Selector," 2020. http://rasbt.github.io/mlxtend/user_guide/feature_selection/SequentialFeatureSelector/ (accessed Nov. 21, 2021).

[15] "Welcome to Lazy Predict's documentation!," 2020. https://lazypredict.readthedocs.io/en/latest/ (accessed Nov. 21, 2021).

[16] Python API Reference, "xgboost," 2021. https://xgboost.readthedocs.io/en/latest/python/python_api.html (accessed Nov. 21, 2021).

[17] Python API, "lightgbm.LGBMClassifier," 2021. https://lightgbm.readthedocs.io/en/latest/pythonapi/lightgbm.LGBMClassifier.html (accessed Nov. 21, 2021).

[18] API Python - scikit learn, "sklearn.ensemble.RandomForestClassifier," 2021. https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html (accessed Nov. 21, 2021).

[19] François Chollet, "Transfer learning & fine-tuning," May 12, 2020. https://keras.io/guides/transfer_learning/ (accessed Nov. 27, 2021).