ANIMUS BEYFORD TRADING

STENING®

Av. Alfredo Palma, 24 - Locales 4 y 5
29603 - Marbella - Málaga - España
CIF B-9348488-9

CSQ MED ISO 13485

Clasificación de dispositivo médico utilizando diferentes técnicas de Machine Learning

Agustín Bignú

Abril 2019

Resumen

En este *paper* se realizó una clasificación de tres clases de dispositivos médicos fabricados por Stening®. La clasificación fue realizada utilizando tres algoritmos de *machine learning* orientados a clasificación. Los algoritmos fueron *Support Vector Machines*, *Logistic Regression* y *Decision Tree*. Con este estudio Stening® pretende dar a conocer más sus productos y que se pueda tener una visión más técnica de estos.

Abstract

In this paper, a classification was made of three classes of medical devices manufactured by Stening®. The classification was made using three machine learning algorithms. These algorithms were Support Vector Machines, Logistic Regression and Decision Tree. With this study Stening® aims to make their products more known and that one can have a more technical vision of these.

Av. Alfredo Palma, 24 - Locales 4 y 5
29603 - Marbella - Málaga - España
CIF B-9348488-9

1. Introducción

Los objetivos principales de este estudio son:

- dar a conocer los productos que Stening® fabrica
- ofrecer una visión más técnica sobre estos utilizando tecnología innovador

En este escrito se realizó una clasificación de tres clases de dispositivos médicos fabricados por Stening®. Para ello se utilizaron tres técnicas diferentes de clasificación de *machine learning*. Se trata de tres algoritmos que hacen lo mismo pero que su funcionamiento interno es totalmente diferente. Estos fueron: Support Vector Machines (SVM), Logistic Regression (LG) y Decision Tree (DT). El funcionamiento interno se explicará en la sección 2.

Los dispositivos médicos que se clasificaron fueron los siguientes:

- ST (Stent Traqueal)
- TM (Tubo en T)
- SY (Stent en "Y")

Se generaron tres datasets de 900, 1200 y 1500 filas para realizar el entrenamiento. Para la clasificación nos basamos en 5 atributos diferentes pero comunes de los dispositivos: largo, diámetro, anclajes, nº de ramas y anchura de la pared. Estos 5 atributos son los que nos diferencian dentro del dataset las tres clases de dispositivo médico. La creación del dataset se explicará en la sección 3.1.

Analizaremos los resultados de cada algoritmo aplicado sobre cada uno de los datasets, esto lo haremos en la sección 3.

Por último, se darán las conclusiones y las vistas a futuro de este estudio.

2. Fundamento teórico

En esta sección introduciremos los algoritmos utilizados para la clasificación de los dispositivos médicos.

Av. Alfredo Palma, 24 - Locales 4 y 5
29603 - Marbella - Málaga - España
CIF B-9348488-9

2.1. SVM

En esta sección se introducirá el algoritmo Support Vector Machines [1]. Se trata de un algoritmo de clasificación perteneciente a la rama de aprendizaje supervisado. La clasificación la realiza encontrando el mejor *hiperplano*[1] en un espacio N-dimensional (*N* es el número de parámetros) que separa a los datos.

Para separar dos clases de datos hay muchos hiperplanos posibles. El objetivo del algoritmo es encontrar el que esté a mayor distancia entre los datos de ambas clases (ver figura 1).

Los hiperplanos marcan la zona donde comienza una clase y acaba otra. Es decir, si en el futuro recibimos un parámetro en la zona baja del hiperplano será de la clase roja, si está en la otra zona será azul. Los hiperplanos pueden ser de diferentes dimensiones, depende de los datos que tengamos. Si tenemos un conjunto de datos $\{x_1, x_2\}$, como en la figura 1, nuestro hiperplano será una recta, si los datos dependen de tres atributos diferentes: $\{x_1, x_2, x_3\}$ el hiperplano será de dimensión dos. En otras palabras, siempre tendrá una dimensión menos que nuestro conjunto de datos.

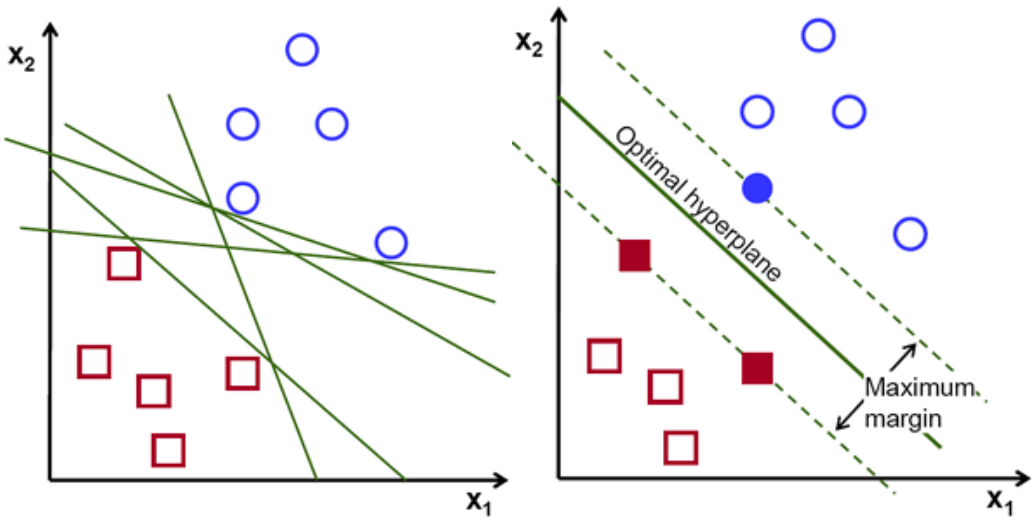


Figura 1: Representación de hiperplanos en dos dimensiones. [2]

Los vectores de apoyo o *support vectors* son datos que están cercanos al hiperplano e influencian su orientación y posicionamiento. En SVM buscamos maximizar el margen entre los puntos y el hiperplano.

La función que nos ayudará a maximizar el margen es la siguiente:

[1] Se trata de un plano de dimensión N-1, siendo N la dimensión total del espacio en el que estemos (1D, 2D, 3D…)

$$c(x, y, f(x)) = \begin{cases} 0, & si\ y * f(x) \geq 1 \\ 1 - y * f(x), & en\ otro\ caso \end{cases} \quad (2.1.1)$$

La expresión (2.1.1) recibe el nombre de *función de pérdida* (loss function). En esta expresión, x son los datos, y es el resultado conocido y $f(x)$ es la predicción que hacemos. Si estas dos últimas son del mismo signo, la función de pérdida vale cero.

Este algoritmo se utiliza para realizar problemas de clasificación y para realizar predicciones. Por ejemplo, distinguir entre tres tipos de flores a partir de tres datos diferentes: $x = \{\text{color, ancho de los pétalos, altura}\}$. Luego, a partir de aprender de los datos, el modelo debe ser capaz de predecir a qué clase de flor (y) pertenece una flor con datos que no haya visto.

2.2. Logistic Regression

La regresión logística es un método estadístico para realizar clasificaciones [3]. Es un tipo especial de regresión lineal dónde las clases a predecir son categóricas. Esto se puede entender con un ejemplo: predecir si tengo una enfermedad o no, las respuestas serían Sí o No.

Como se dijo es un tipo especial de regresión lineal. La fórmula de una regresión lineal es la siguiente:

$$y = \beta_0 + \beta_1 X_1 + \cdots + \beta_n X_n \quad (2.2.1)$$

En la expresión (2.2.1) y es el resultado de la predicción y X_1, X_2, ... son las variables con las que entrena el modelo. La regresión logística la obtenemos introduciendo (2.2.1) en la siguiente expresión:

$$p = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \cdots + \beta_n X_n)}} \quad (2.2.2)$$

La expresión (2.2.2) recibe el nombre de función Sigmoidea (figura 2). Entonces, la regresión logística consiste en aplicar la función sigmoidea a una regresión lineal.

Tenemos tres tipos de regresiones logísticas:

- Regresión logística binaria: dos categorías para predecir
- Regresión logística multinomial: tres o más categorías para predecir
- Regresión logística ordinal: tres o más categorías para predecir pero ordinales, es decir, con cierto orden.

En este escrito utilizaremos la regresión logística multinomial ya que tenemos tres clases de dispositivos médicos, no será ordinal porque carecen de orden.

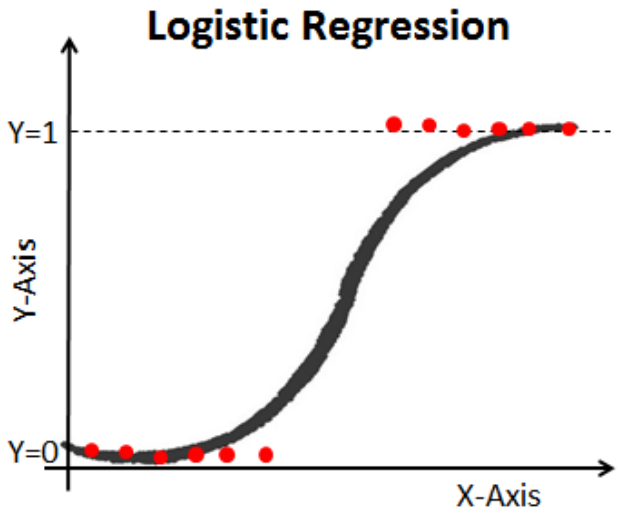


Figura 2: Representación gráfica de una regresión logística binaria. [3]

2.3. Decision Tree

Un árbol de decisión es un algoritmo ampliamente utilizado, no sólo en Machine Learning sino en otras áreas de la computación [4]. Su lógica resulta más simple que la de los otros dos algoritmos ya que es más intuitiva. Esta lógica la explicaremos a continuación.

El árbol de decisión tiene tres componentes principales: nodos, hojas y ramas. Un nodo representa un atributo, una rama representa una decisión y una hoja (*leaf*) representa una salida. El objetivo principal es generar un árbol de decisión que tenga tantas hojas como categorías quieras clasificar, en nuestro caso tres ya que tenemos tres tipos diferentes de dispositivo médico. La estructura sería como la de la figura 3.

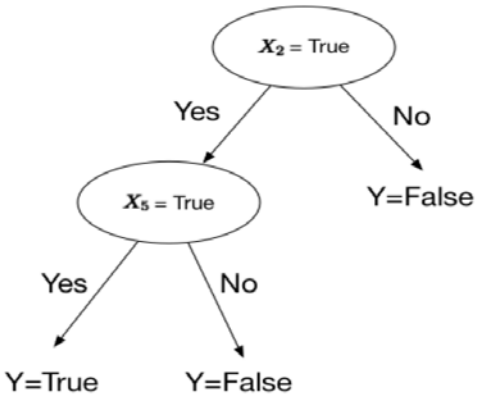


Figura 3: Representación gráfica de un árbol de decisión. [5]

Av. Alfredo Palma, 24 - Locales 4 y 5
29603 - Marbella - Málaga - España
CIF B-9348488-9

3. Resultados

En esta sección vamos a explicar cómo realizamos el dataset y los resultados obtenidos.

3.1. Dataset

El dataset lo realizamos a partir de las medidas de los dispositivos originales de Stening®. Se realizaron tres datasets diferentes de 900, 1200 y 1500 filas cada uno.

Como se dijo, a partir de las dimensiones de los dispositivos fabricados por Stening®, fuimos generando los datasets aleatoriamente. Esto nos dio una libertad de entrenamiento para tener una mayor variedad. Esto es así porque algunas dimensiones que se incluyen en su página web no son las más vendidas y por tanto las menos fabricadas, por lo que si nos basamos en eso para generar el dataset, por más que estuviera más cercano a la realidad estaría sesgando el modelo y el entrenamiento. Es por eso que, como primera aproximación, decidimos generar el dataset de esta forma. No obstante, como se dijo, siempre dentro de las dimensiones reales de los dispositivos.

En la siguiente imagen podemos ver una serie de filas del dataset de 1200:

	Largo	Diametro	Anclajes	Numero Ramas	Ancho pared	Clases
0	80.0	15.0	1.0	0.0	1.5	0.0
1	70.0	5.0	1.0	0.0	0.5	0.0
2	60.0	18.0	1.0	0.0	1.5	0.0
3	80.0	6.0	1.0	0.0	0.6	0.0
4	40.0	12.0	1.0	0.0	1.2	0.0
5	60.0	10.0	1.0	0.0	1.0	0.0
6	40.0	10.0	1.0	0.0	1.0	0.0
7	40.0	15.0	1.0	0.0	1.5	0.0
8	30.0	13.0	1.0	0.0	1.5	0.0
9	40.0	15.0	1.0	0.0	1.5	0.0
10	50.0	14.0	1.0	0.0	1.5	0.0

Figura 4: Primeras 11 filas del dataset de 1200 filas.

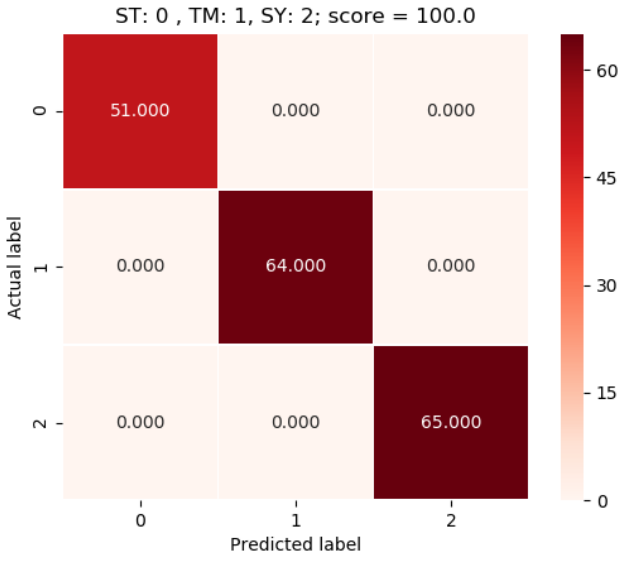
Al ser 3 clases diferentes de dispositivo, se les asigno un número para poder entrenar los modelos. De esta forma: '0' representa un ST, '1' un TM y '2' un SY. A su vez, para los parámetros binarios (Si/No) como la columna ("Anclajes"), se le asignó un '1' si la respuesta era 'Sí' y un '0' si la respuesta era 'No'.

Para la realización del dataset se utilizó un programa escrito en Python 3.5.

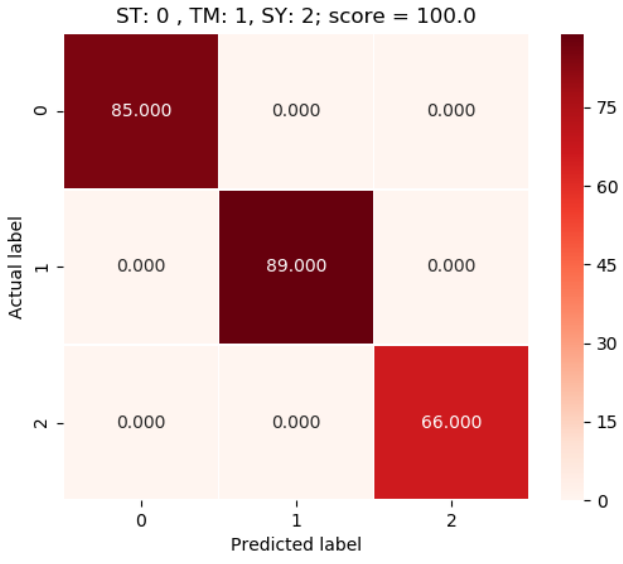
3.2. Programas y resultados

Se realizaron tres programas, uno para cada modelo a aplicar. A estos tres programas se les introdujo cada uno de los tres datasets. Obtuviendo así, tres resultados por algoritmo. Estos tres programas se escribieron, como el que genera el dataset, en Python 3.5. Para realizarlos se utilizó la librería de machine learning *Sklearn*.

En cuanto a los resultados obtenidos, se obtuvieron eficacia absoluta en los tres modelos. Cada uno de ellos fue capaz de obtener una eficacia del 100% en la predicción de muestras que no había visto antes. En las imágenes siguientes podemos ver los resultados obtenidos:

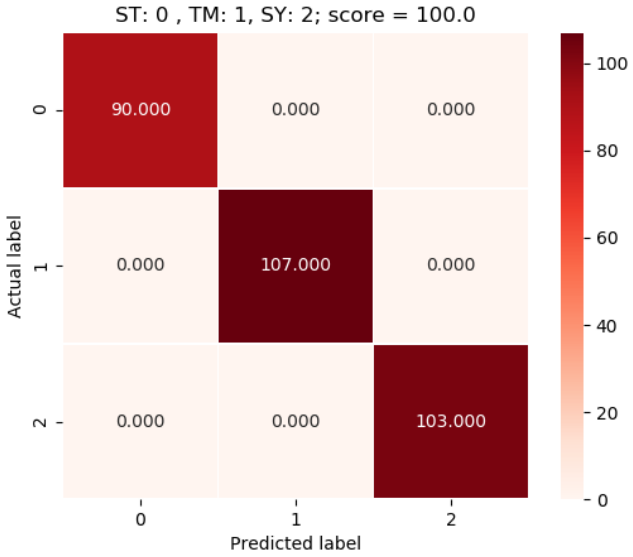


A

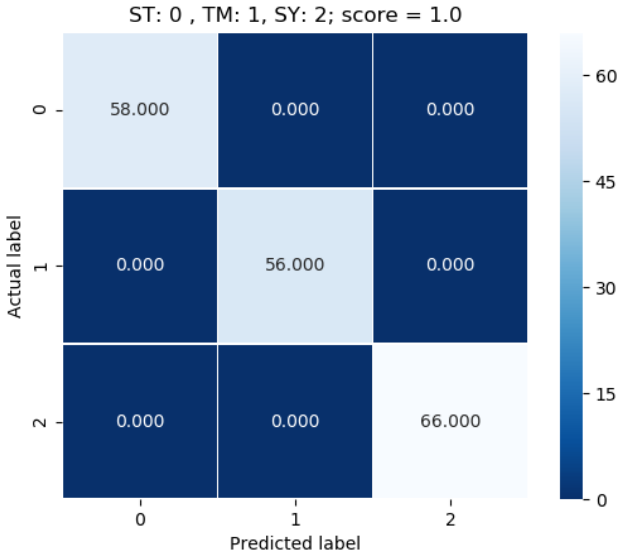


B

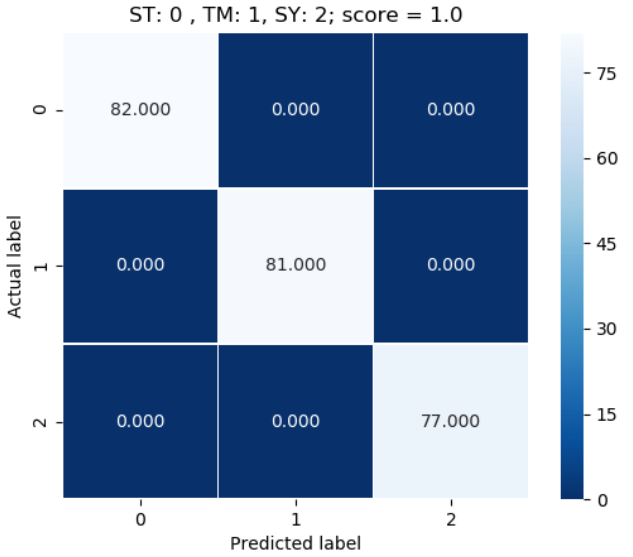
Av. Alfredo Palma, 24 - Locales 4 y 5
29603 - Marbella - Málaga - España
CIF B-9348488-9



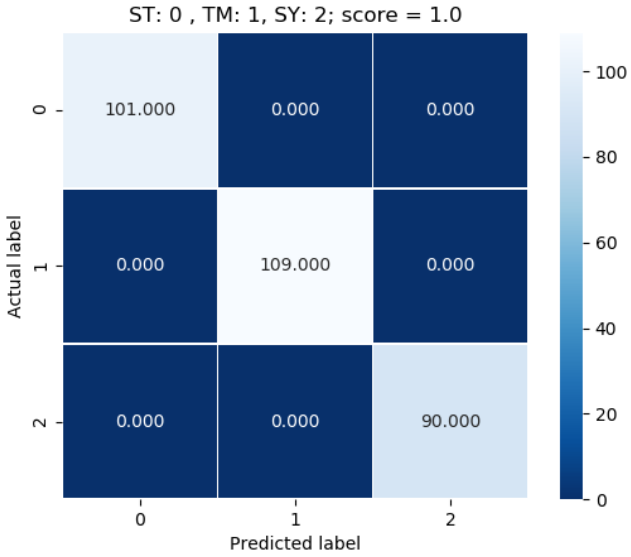
D



E

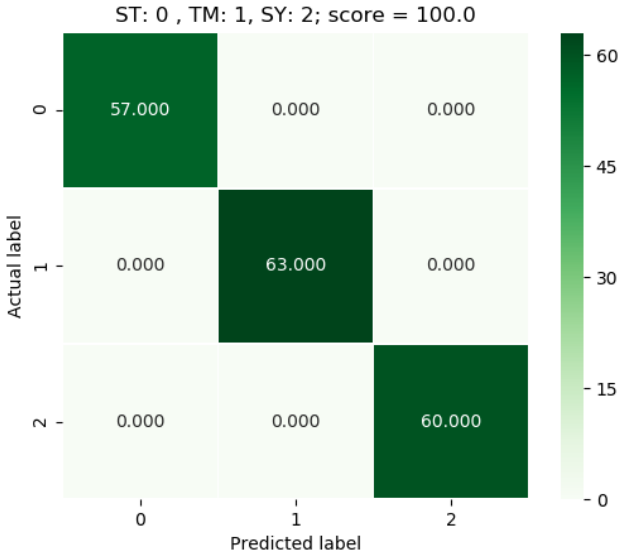


F

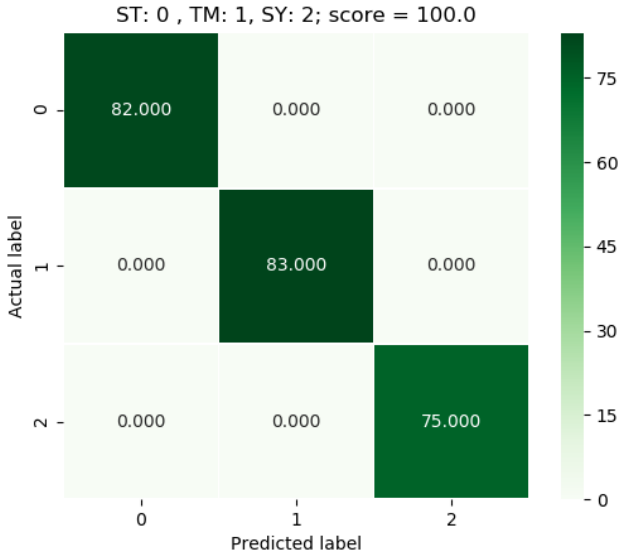


G

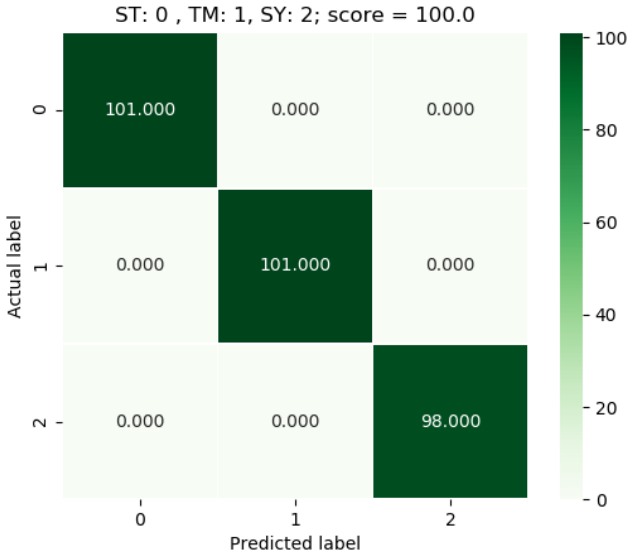
Av. Alfredo Palma, 24 - Locales 4 y 5
29603 - Marbella - Málaga - España
CIF B-9348488-9



H



I



J

Figura 5: verde (SVM), azul (Logistic Regression) y rojo (Decision Tree).

En la figura 5 podemos ver los resultados obtenidos para cada dataset. Como se dijo, hay tres cuadros de resultados por algoritmo. Están en orden, siendo el primero el de 900 y el último el de 1500. En cada cuadro podemos ver una referencia a las labels. En el eje vertical tenemos la label correcta y en el horizontal la label predicha por el modelo. En la esquina superior derecha tenemos el resultados de la predicción. El número que aparece dentro del cuadrado es el número de dispositivos predichos con ese

label. Si sumamos todos los números veremos que no se obtiene el total del dataset ya que estas son las predicciones sobre una parte del dataset que el modelo no vio durante el entrenamiento. Esto se debe a que el dataset se dividió en 70% para entrenamiento y 30% para realizar predicciones.

4. Conclusiones

Para concluir cabe mencionar que los resultados obtenidos son muy satisfactorios. Esto nos motiva a seguir investigando y realizando estudios relacionados con el machine learning y la inteligencia artificial.

En futuros estudios se harán mejoras como entrenar con dispositivos reales en stock de mercado. Así como aplicar técnicas similares de machine learning a otras ramas de dispositivos de Stening®.

Referencias

[1] Support Vector Machine, https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47

[2] Support Vector Machine, https://towardsdatascience.com/ support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47 Fig. 1

[3] Logistic Regression, https://www.datacamp.com/community/tutorials/understanding-logistic-regression-python

[4] Decision Tree, https://medium.com/deep-math-machine-learning-ai/chapter-4-decision-trees-algorithms-b93975f7a1f1

[5] Decision Tree, https://hackernoon.com/what-is-a-decision-tree-in-machine-learning-15ce51dc445d