



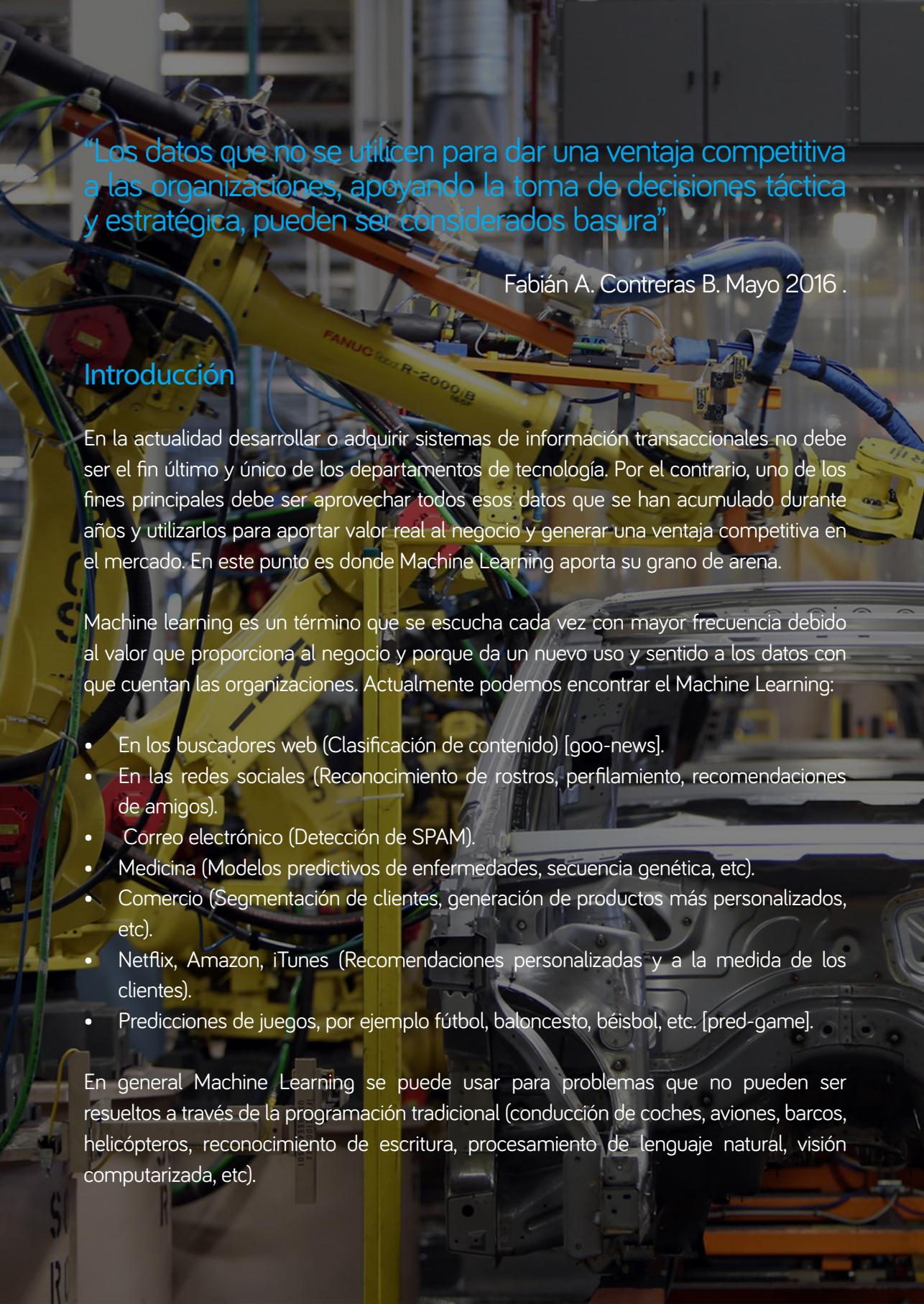
 SUNQU

# INTRODUCCIÓN A MACHINE LEARNING

por Fabián A. Contreras, Arquitecto de Soluciones

## Índice

Introducción	3
Definición	5
Métodos de Machine Learning	6
Método de Regresión	6
Método de Clasificación	7
Método de Agrupación	8
Tipos de Aprendizaje Automático	9
Aprendizaje Supervisado	9
Aprendizaje No Supervisado	10
Representación del Modelo Matemático de Machine Learning	11
Regularización	13
Sobreajuste - Alta Varianza (Overfitting)	13
Subajuste - Alta Oscilación (Underfitting)	13
Regularización	14
Modelos de Aplicación	15
Consejos de mejoras a los modelos de ML	16
Metodología aplicación ML	18
Arquitectura Conceptual ML	21
WSO2 Machine Learner	22
Casos prácticos	23
Referencias	24



“Los datos que no se utilicen para dar una ventaja competitiva a las organizaciones, apoyando la toma de decisiones táctica y estratégica, pueden ser considerados basura”.

Fabián A. Contreras B. Mayo 2016 .

## Introducción

En la actualidad desarrollar o adquirir sistemas de información transaccionales no debe ser el fin último y único de los departamentos de tecnología. Por el contrario, uno de los fines principales debe ser aprovechar todos esos datos que se han acumulado durante años y utilizarlos para aportar valor real al negocio y generar una ventaja competitiva en el mercado. En este punto es donde Machine Learning aporta su grano de arena.

Machine learning es un término que se escucha cada vez con mayor frecuencia debido al valor que proporciona al negocio y porque da un nuevo uso y sentido a los datos con que cuentan las organizaciones. Actualmente podemos encontrar el Machine Learning:

- En los buscadores web (Clasificación de contenido) [goo-news].
- En las redes sociales (Reconocimiento de rostros, perfilamiento, recomendaciones de amigos).
- Correo electrónico (Detección de SPAM).
- Medicina (Modelos predictivos de enfermedades, secuencia genética, etc).
- Comercio (Segmentación de clientes, generación de productos más personalizados, etc).
- Netflix, Amazon, iTunes (Recomendaciones personalizadas y a la medida de los clientes).
- Predicciones de juegos, por ejemplo fútbol, baloncesto, béisbol, etc. [pred-game].

En general Machine Learning se puede usar para problemas que no pueden ser resueltos a través de la programación tradicional (conducción de coches, aviones, barcos, helicópteros, reconocimiento de escritura, procesamiento de lenguaje natural, visión computarizada, etc).



Aunque el Machine Learning no es un concepto nuevo, en los últimos años ha tenido un auge exponencial en su uso y aplicación. Una de las razones principales de este auge es el aumento en la capacidad de procesamiento y la disminución de los costes del mismo, permitiendo así que pueda estar al alcance de todos.

El objetivo de este documento es presentar los conceptos fundamentales vinculados con el Machine Learning. También se mostrará la manera en que WSO2 ML implementa estos conceptos de manera sencilla con el fin de realizar aplicaciones prácticas.

El contenido teórico de este artículo está basado en las notas del curso de Machine Learning de la Universidad de Stanford, dictado por el profesor asociado Andrew Ng [and-ml], ofrecido por Coursera [cou-ml].

## 2. Definición

El Machine Learning es una rama de la inteligencia artificial encargada de crear programas de software capaces de generalizar comportamientos a partir de los datos recibidos [wik-ml]. Otros autores definen este concepto de la siguiente manera:

- Enseñar a un computador a aprender conceptos usando datos, sin ser explícitamente programado para ello.
- Campo de estudio que da a los ordenadores la habilidad de aprender sin la necesidad de ser explícitamente programados [Arthur Samuel, 1959].
- Se dice que un programa de ordenador aprende por medio de la experiencia E con respecto a alguna clase de tareas T y medida de rendimiento P, si su desempeño en tareas en T, medida por P, mejora con la experiencia E. [Tom Mitchel, 1998].

Ejemplo: Jugando al ajedrez.

E = La experiencia de jugar muchos juegos de ajedrez.

T = La tarea de jugar ajedrez.

P = La probabilidad que el programa gane el próximo partido.



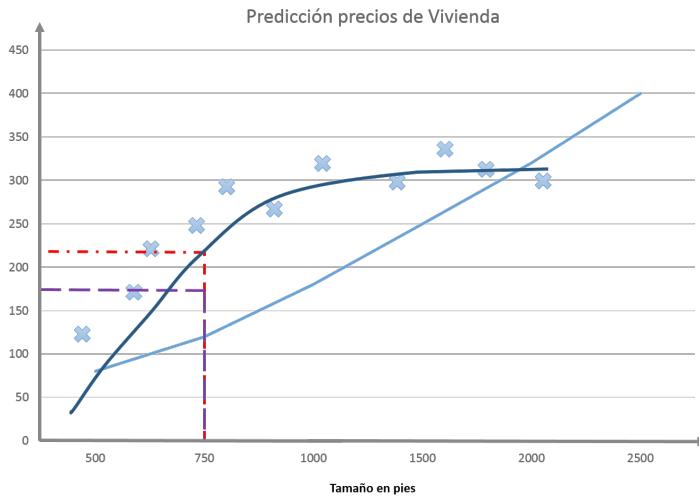
### 3. Métodos de Machine Learning

Dentro del Machine Learning existen tres métodos de aplicación diferenciados, los cuales se explicarán a continuación de manera específica.

#### 3.1 Método de Regresión.

Este método se utiliza para predecir el valor de un atributo continuo. Consiste en encontrar la mejor ecuación que atraviese de forma óptima un conjunto de puntos ( $n$ -dimensiones). Se utiliza cuando la precisión no es crítica y el número de variables es pequeño.

Ej: Predecir el precio de una vivienda, dado su tamaño.



Ejemplo de modelo  
regresión lineal.

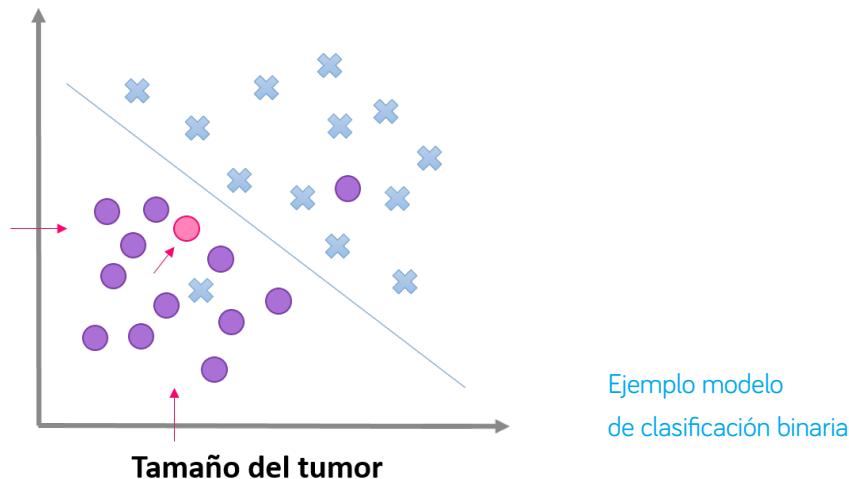


### 3.2 Método de Clasificación.

Método utilizado para predecir un resultado de un atributo con valor discreto (a, b, c, ...) dadas unas características ( $X_0, X_1, X_2, X_3, \dots, X_n$ ).

El método simple de clasificación es el binario, donde se clasifica un registro de variables de entrada en 1 o 0. La clasificación múltiple es una extensión de la clasificación binaria.

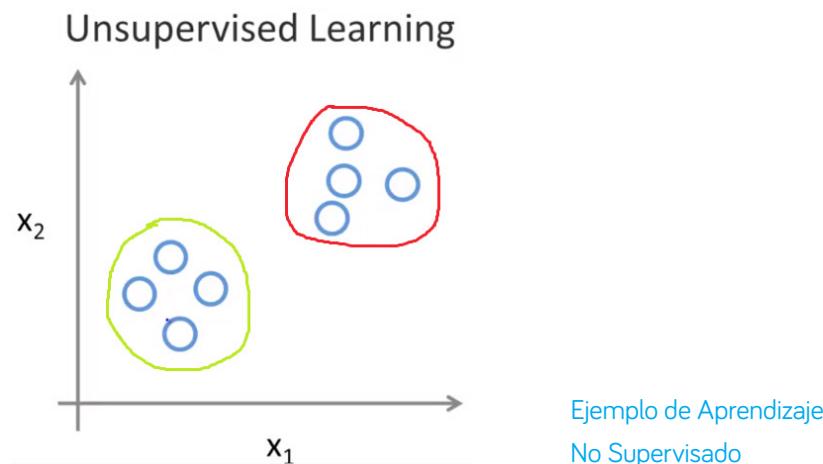
Ej: Identificar si un tumor es maligno o benigno, dado su tamaño y edad del paciente.



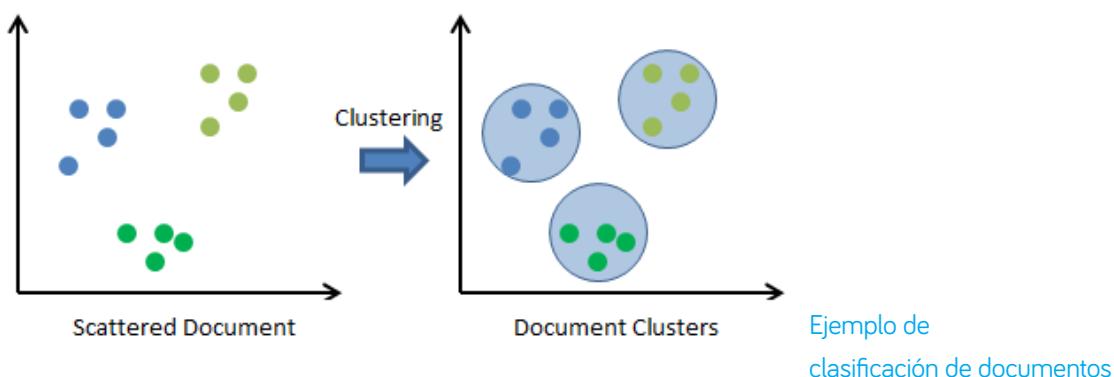
### 3.3. Método de Agrupación.

Este método se utiliza cuando se necesita clasificar las instancias de datos pero no se conocen previamente las categorías. Esta agrupación permite construir grupos (cluster) coherentes de instancias teniendo en cuenta las variables de la data. En palabras sencillas, permite encontrar qué se tiene en la data.

Ej: Falta ejemplo de Cluster.



Ej: Ejemplo de Clasificación de Documentos. No se sabe de antemano que tipos de documentos se tiene, al aplicar el algoritmo de clasificación se encuentra que se tienen 3 clasificadores del conjunto de documentos.



## 4. Tipos de Aprendizaje Automático

Dentro del Machine Learning los algoritmos utilizados los podemos agrupar en dos tipos:

### 4.1 Aprendizaje Supervisado.

El aprendizaje supervisado es aquel que para un conjunto de datos de entrada conocemos de antemano los datos correctos de salida. Consta de 2 fases, una de entrenamiento y otra de pruebas.

- En la fase de entrenamiento se cuenta con un conjunto de datos (por lo general entre el 60% o 70% del total de la data disponible) que son con los que se enseña o entrena al algoritmo para encontrar los patrones y relaciones en la data.
- Posteriormente en la fase de pruebas se cuenta con una data de pruebas (entre el 40% o 30% del total de la data disponible), la cual sirve para validar el rendimiento del algoritmo.

Los métodos enunciados anteriormente que se encuentran en este tipo de algoritmos son:

- Regresión: se trata de predecir resultados con una salida continua. Es decir, mapear variables de entrada a alguna función continua.
- Clasificación: se trata de predecir resultados con una salida discreta. O lo que es lo mismo, mapear variables de entrada en categorías discretas.

Ej:

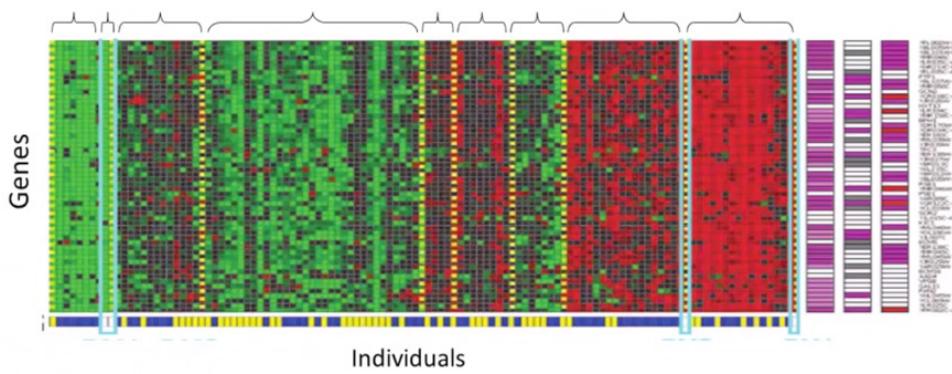
- Reconocimiento de Voz. [spe-rec]
- Detección de fraude en tarjetas de crédito. [fra-det]

## 4.2 Aprendizaje No Supervisado.

En el aprendizaje no supervisado para un conjunto de datos de entrada, NO conocemos de antemano los datos de salida. Se utiliza para obtener una agrupación coherente de los datos en función de las relaciones entre las variables definidas en la data.

Ej:

- Tome una colección de 1000 ensayos escritos sobre la economía de Colombia, y encuentre una manera de forma automática de clasificar estos ensayos en grupos, que son de alguna manera similares o relacionados por diferentes variables, como la frecuencia de la palabra, longitud de la oración, número de páginas, y así sucesivamente. (Google News).
- Clasificación de personas dados sus genomas.



Ejemplo de Clasificación de Genoma Humano

## 5. Representación del Modelo Matemático de Machine Learning

Lograr que las máquinas aprendan comportamientos, encuentren patrones o realicen predicciones a partir de datos no es magia. Se puede lograr con la aplicación de las matemáticas y, en el caso particular del Machine Learning, mediante operaciones entre matrices (Algebra Lineal).

A continuación se presenta una descripción sencilla del modelo matemático que existe detrás de los algoritmos utilizados en Machine Learning.

- Definir una hipótesis  $h(\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3 \dots \mathbf{X}_n)$  en función de las características o variables que se quieran manejar (estas características deben ser numéricas).

$$h_{\theta}(\mathbf{x}) = \theta_0 + \theta_1 x$$

**x (input) y (output)**

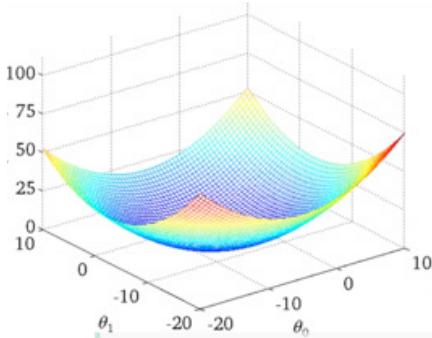
0	4
1	7
2	7
3	8

$$\theta_0 = 2 \text{ and } \theta_1 = 2.$$

$$h_{\theta}(\mathbf{x}) = 2 + 2x.$$

Ejemplo de Clasificación  
de Genoma Humano

b. Definir una función de costos alrededor de la hipótesis,  $J(T_1, T_2, T_3 \dots T_n)$ . Esta función debe ser Convexa. Es decir, que tenga un único mínimo global.



### Cost function

For the parameter vector  $\theta$  (of type  $\mathbb{R}^{n+1}$  or in  $\mathbb{R}^{(n+1) \times 1}$ )

$$J(\theta) = \frac{1}{2m} \sum_{i=1}^m (h_\theta(x^{(i)}) - y^{(i)})^2$$

The vectorized version is:

$$J(\theta) = \frac{1}{2m} (X\theta - \vec{y})^T (X\theta - \vec{y})$$

Where  $\vec{y}$  denotes the vector of all y values.

### Ejemplo de Función de Costos Convexa

c. Definir un algoritmo para minimizar la función de costos (encontrar la derivada de la función de costos que encuentre la combinación de  $T_1, T_2, T_3 \dots T_n$  que obtenga el valor mínimo de  $J$ ).

The gradient descent equation itself is generally the same form

repeat until convergence: {

$$\theta_0 := \theta_0 - \alpha \frac{1}{m} \sum_{i=1}^m (h_\theta(x^{(i)}) - y^{(i)}) \cdot x_0^{(i)}$$

$$\theta_1 := \theta_1 - \alpha \frac{1}{m} \sum_{i=1}^m (h_\theta(x^{(i)}) - y^{(i)}) \cdot x_1^{(i)}$$

$$\theta_2 := \theta_2 - \alpha \frac{1}{m} \sum_{i=1}^m (h_\theta(x^{(i)}) - y^{(i)}) \cdot x_2^{(i)}$$

...

}

In other words:

repeat until convergence: {

$$\theta_j := \theta_j - \alpha \frac{1}{m} \sum_{i=1}^m (h_\theta(x^{(i)}) - y^{(i)}) \cdot x_j^{(i)} \quad \text{for } j := 0..n$$

}

**Algoritmo Gradiante  
Descendiente**

## 6. Regularización

Para entender este término, primero debemos definir dos conceptos claves que se manejan en Machine Learning y que nos sirven para entender como se está comportando el algoritmo: Sobreajuste (Overfitting) y Subajuste (UnderFiting).

### 6.1 Sobreajuste - Alta Varianza (Overfitting).

La hipótesis o función seleccionada mapea casi perfectamente la tendencia de los datos de entrenamiento, pero puede fallar en la generalización de nuevos registros.

Para combatir el sobreajuste tenemos las siguientes opciones:

- Reducir el número de características o eliminar las no relevantes.
- Seleccionar otro tipo de algoritmo que maneje mejor este problema.
- Aplicar la Regularización.

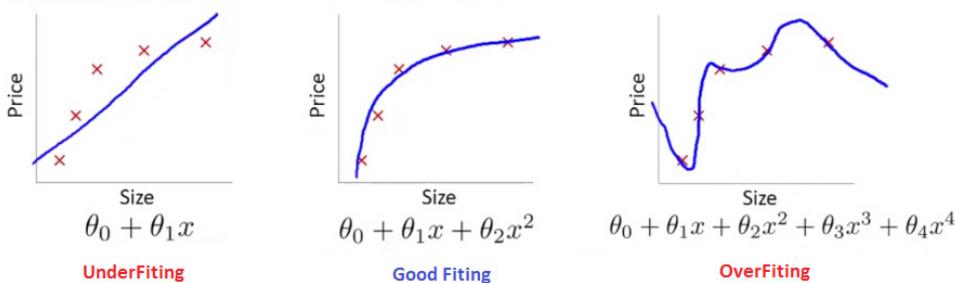
### 6.2 Subajuste - Alta Oscilación (Underfitting).

La hipótesis o función seleccionada mapea pobemente la tendencia de los datos de entrenamiento, con lo cual también tiene problemas con la generalización de nuevos registros.

Para combatir el subajuste tenemos las siguientes opciones:

- Aumentar el número de características relevantes.
- Incluir un número mayor de registros.

**Linear regression (housing prices)**



Diferentes formas de Ajuste

### 6.3 Regularización.

La regularización es una técnica que se utiliza para combatir el sobreajuste. La regularización funciona bien cuando tenemos una gran cantidad de características poco útiles.

Este método consiste en obtener valores pequeños para los parámetros T. De esta forma se obtiene una hipótesis más sencilla y se previene el sobreajuste. Para ellos se incluye un término adicional a la función de costos que es la sumatoria de los valores de T al cuadrado por un factor de regularización lambda.

La función de costos es  $J(T) = \frac{1}{2m} [\sum ((h(x(i)) - y(i))^2 + \lambda \sum \theta_j^2)]$

$$\min_{\theta} \frac{1}{2m} \left[ \sum_{i=1}^m (h_{\theta}(x^{(i)}) - y^{(i)})^2 + \lambda \sum_{j=1}^n \theta_j^2 \right]$$

Función de Costos con Regularización

El primer término de la función de costos es el encargado de que la hipótesis se ajuste bien a los puntos de entrenamiento; el segundo se encarga de disminuir los valores de los parámetros T; lambda es el término que compensa el objetivo de ambas metas.

Es necesario tener en cuenta que:

- Lambda grande  $\rightarrow$  subajuste. Debido a que penalizamos a los parámetros T. Lo cual haría que obtuviéramos una hipótesis  $h(x) = T_0$ .
- Lambda pequeño  $\rightarrow$  sobreajuste. Debido a que el aporte en el término de regularización es despreciable.

## 7. Modelos de Aplicación

Un modelo es un algoritmo de Machine Learning (en adelante ML) aplicado a unos datos para resolver un problema en particular [pri-ml]. Dentro de los diferentes algoritmos de ML que podemos utilizar para la generación de nuestros modelos se encuentran:

- Regresión Lineal (Predecir un valor).
- Regresión Logística (Predecir una Clasificación).
- Redes Neuronales (Predecir una Clasificación).
- Support Vector Machine (SVM). (Predecir una Clasificación). De los más utilizados hoy en día.
- K-means (Clustering – Agrupamiento).
- Algoritmo de Distribución Gaussiana (Detección de Anomalías).
- K-means (Detección de Anomalías).
- Algoritmo de filtrado colaborativo (Sistemas de Recomendación).
- Vectorización de bajo rango de matriz de factorización (Sistemas de Recomendación).

## 8. Consejos de mejoras a los modelos de ML

Para mejorar el rendimiento de los Modelos de ML se pueden tener en cuenta las siguientes recomendaciones:

- Obtener más ejemplos de entrenamiento.
- Tratar con un conjunto de características más pequeño.
- Tratar de obtener características adicionales.
- Añadir características polinomiales ( $x_1^2, x_2^2, x_1x_2$ ).
- Ajustar lambda.

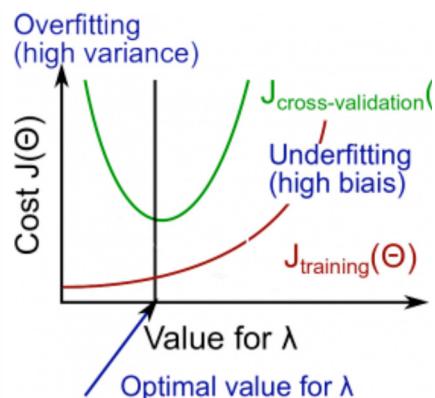
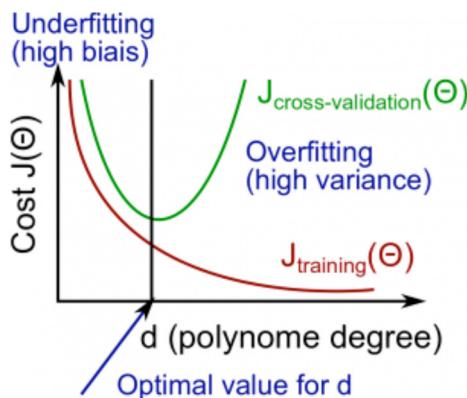
Para aumentar la velocidad de convergencia del Algoritmo, debemos tener en cuenta lo siguiente:

- Podemos utilizar dos técnicas para escalar los valores de las características (que todas las características tengan la misma escala de valores):
- Los rangos más utilizados son  $-1 \leq x \leq 1$  ó  $-0.5 \leq x \leq 0.5$
- Escalar las características.  $x_s = (x_{(i)j}) / (x_{\max} - x_{\min})$ .
- Normalización media.  $x_s = (x_{(i)j} - u_i) / (x_{\max} - x_{\min})$ .

Sin embargo no se debe escoger alguna o algunas al azar, ya que existen técnicas que ayudan a seleccionar qué recomendaciones aplicar. Son las siguientes:

Evalué la Hipótesis seleccionada.

- Prepare un conjunto de datos de validación (diferentes a los de entrenamiento, al menos el 20% del total de registros) para validar el TEST de Error. Con este conjunto de datos realice el ajuste de los parámetros Theta y demás parámetros propios del Algoritmo.
- Prepare un conjunto de datos de prueba (diferentes a los de entrenamiento y validación, al menos el 20% del total de registros) para validar el TEST de Error.
- $\text{Error} = 1/m \left[ \sum (h(x_i) - y_i) \right]$
- Determinar si se tienen problemas de Alto sesgo o Alta Varianza:
- Alto Sesgo (High bias)  $\rightarrow$  Underfitting  $\rightarrow J_{\text{training}} \text{ y } J_{\text{cross-validation}}$  son altos también  $J_{\text{training}}$  es similar  $J_{\text{cross-validation}}$ .
- Alta Varianza (High variance)  $\rightarrow$  Overfitting  $\rightarrow J_{\text{training}}$  es bajo y  $J_{\text{cross-validation}}$   $\gg J_{\text{training}}$ .
- Impacto de la Regularización.
- Lambda grande  $\rightarrow$  Alto Sesgo (High bias)  $\rightarrow$  Underfitting  $\rightarrow J_{\text{training}} \text{ y } J_{\text{cross-validation}}$  son altos.
- Lambda pequeño  $\rightarrow$  Alta Varianza (High variance)  $\rightarrow$  Overfitting  $\rightarrow J_{\text{training}}$  es bajo y  $J_{\text{cross-validation}}$  es alto.

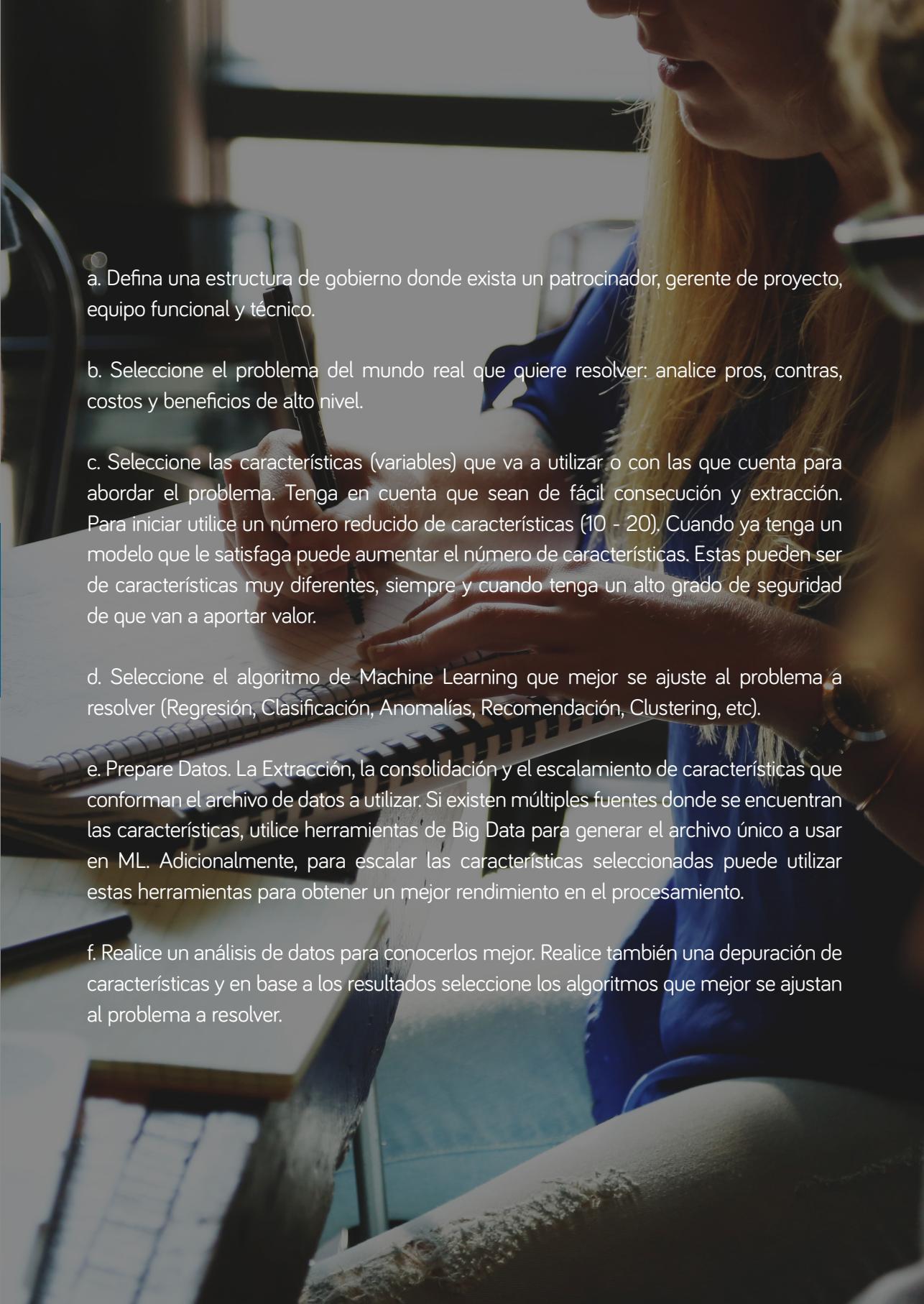


Verificaciones del rendimiento del Modelo

Es importante resaltar que el resultado del rendimiento de un algoritmo es bueno o malo dependiendo del contexto de negocio en el que se quiera aplicar. Por ejemplo, si se trabaja en un modelo predictivo de riesgo de Cáncer y lo queremos utilizar para informar al paciente de la probabilidad de contraer la enfermedad, podríamos informarle de que es necesaria una precisión y exactitud de más del 90% para evitar impactos emocionales en el paciente. Pero si por el contrario necesitamos la predicción para incluir al paciente en algún nivel de los programas de prevención y promoción tal vez requiera una precisión y exactitud de entre 50% y 60%.

## 9. Metodología aplicación ML



- 
- a. Defina una estructura de gobierno donde exista un patrocinador, gerente de proyecto, equipo funcional y técnico.
  - b. Seleccione el problema del mundo real que quiere resolver: analice pros, contras, costos y beneficios de alto nivel.
  - c. Seleccione las características (variables) que va a utilizar o con las que cuenta para abordar el problema. Tenga en cuenta que sean de fácil consecución y extracción. Para iniciar utilice un número reducido de características (10 - 20). Cuando ya tenga un modelo que le satisfaga puede aumentar el número de características. Estas pueden ser de características muy diferentes, siempre y cuando tenga un alto grado de seguridad de que van a aportar valor.
  - d. Seleccione el algoritmo de Machine Learning que mejor se ajuste al problema a resolver (Regresión, Clasificación, Anomalías, Recomendación, Clustering, etc).
  - e. Prepare Datos. La Extracción, la consolidación y el escalamiento de características que conforman el archivo de datos a utilizar. Si existen múltiples fuentes donde se encuentran las características, utilice herramientas de Big Data para generar el archivo único a usar en ML. Adicionalmente, para escalar las características seleccionadas puede utilizar estas herramientas para obtener un mejor rendimiento en el procesamiento.
  - f. Realice un análisis de datos para conocerlos mejor. Realice también una depuración de características y en base a los resultados seleccione los algoritmos que mejor se ajustan al problema a resolver.

g. Realice un análisis del rendimiento de los modelos generados a partir de los algoritmos seleccionados, teniendo en cuenta lo siguiente:

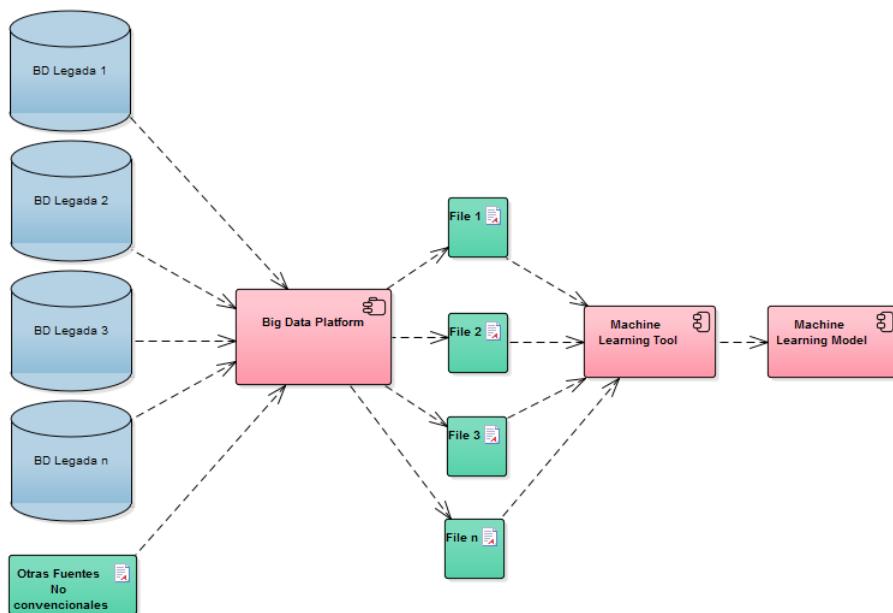
- Utilice el 60% de los datos recolectados como datos de entrenamiento.
- Utilice el 20% de los datos recolectados para hacer una validación cruzada del rendimiento del modelo generado para depurarlo y ajustarlo.
- Utilice el 20% de los datos recolectados para hacer una prueba del modelo generado y comprobar el rendimiento obtenido.

h. Publique el modelo y úselo en un ambiente productivo.

i. Monitoree y ajuste. Con el modelo en producción monitoree su desempeño y haga los ajustes necesarios para mejorar el desempeño cuando sea necesario. Revise las características para incluir más a las actuales y la posibilidad de obtener más datos de entrenamiento, validación y pruebas.

## 10. Arquitectura Conceptual ML.

En las organizaciones actuales existe una gran variedad de repositorios de datos como bases de datos relacionales, archivos planos y excel, entre otros. Estos repositorios almacenan datos estructurados y no estructurados. Adicionalmente se puede contar con datos que están viajando entre aplicaciones en formato XML y/o JSON; para poder sacar provecho de todos estos datos, se sugiere contar con una plataforma que interactúe con estas fuentes, permita su tratamiento y preparación para obtener información que apoye la toma de decisiones tácticas y estratégicas en la organización.



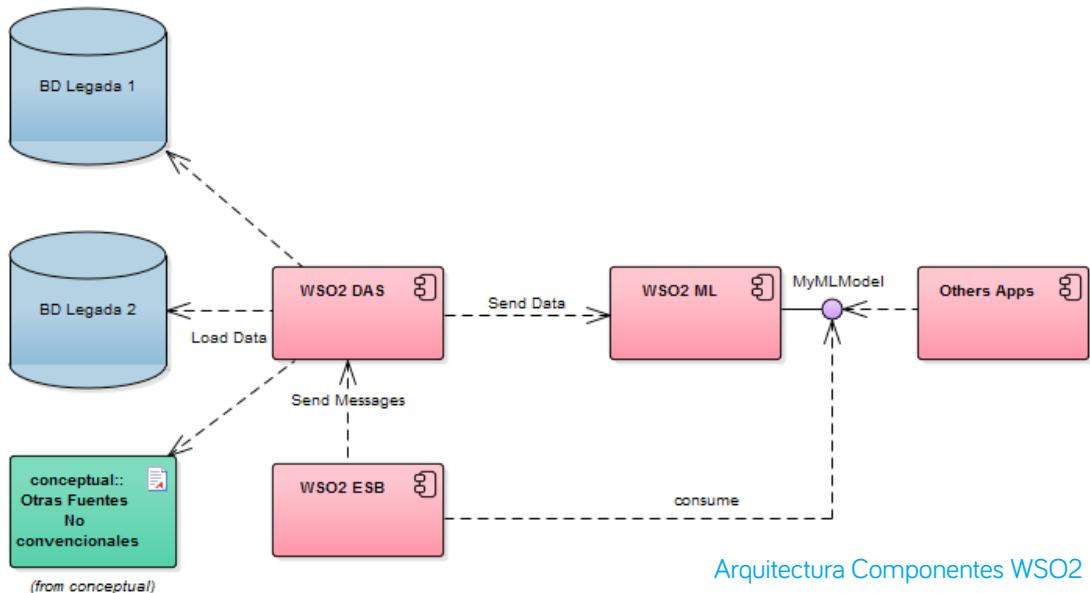
Arquitectura Conceptual ML

## 11. WSO2 Machine Learner

WSO2-ML es una herramienta sencilla y completa para expertos y novatos que estén trabajando en generar modelos predictivos a partir de datos.

Está basada en la librería de ML de Spark, la cual implementa varios algoritmos de regresión, clasificación y agrupamiento. Incorpora la característica de escalado de variables para aumentar la velocidad de convergencia de los algoritmos. Incorpora además un soporte para realizar un análisis visual de los datos con los que se quiere trabajar, permitiendo a los científicos de datos conocer mejor las relaciones existentes en sus datos. Adicionalmente permite configurar el porcentaje de datos para realizar el entrenamiento y las respectivas pruebas, visualizando el rendimiento del modelo basado en distintas métricas estadísticas (ROC curve, Fscore, etc). Por último, una característica diferenciadora es que el modelo predictivo generado se expone como servicios REST, lo cual permite que se pueda interactuar con él desde otros sistemas de información [nir-wso2-ml].

Se propone una arquitectura de analítica basada en la suite de productos de WSO2 [wso2-prod]:



Arquitectura Componentes WSO2

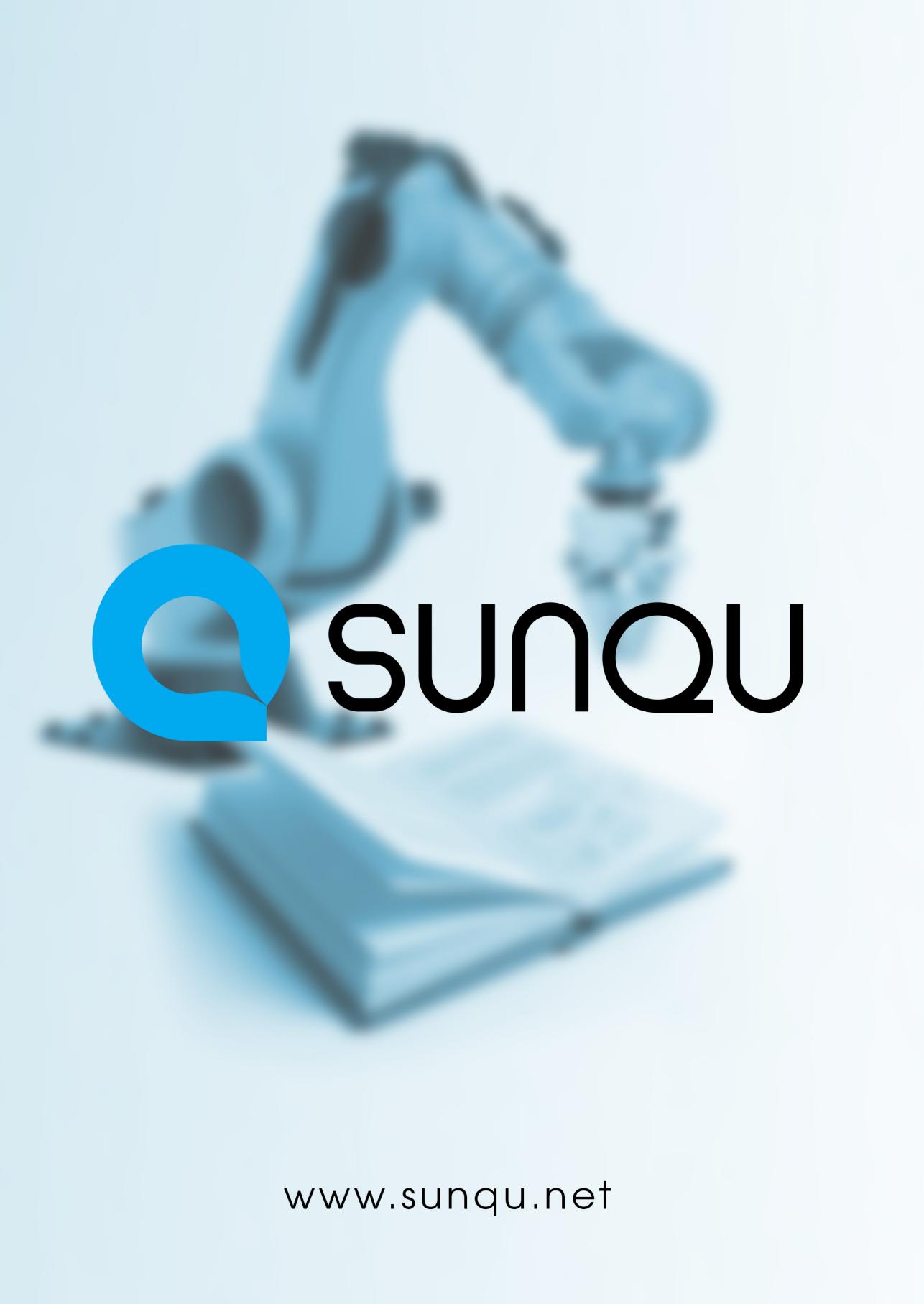
## 12. Casos prácticos

Los anteriores conceptos, metodología y herramientas propuestas, se están utilizando para la implementación de modelos predictivos pilotos en los siguientes temas:

- Modelo de prevención de Morosos. Modelo basado en datos históricos de pagos y valores de cuota, además de datos demográficos y socio-económicos del entorno. El objetivo es clasificar a la persona en una pirámide de riesgo de morosidad, donde cada nivel tiene asociadas unas acciones encaminadas a disminuir oportunamente el riesgo de morosidad.
- Modelo para la prevención del abandono escolar de la educación secundaria, apoyada en modelos predictivos de Machine Learning. Modelo basado en datos del rendimiento académico y datos socio-económicos del entorno familiar. El objetivo es clasificar a los estudiantes en una pirámide de riesgo de abandono escolar, donde cada nivel tiene asociadas unas acciones encaminadas a prevenir y mitigar la deserción escolar.
- Modelo predictivo de riesgo cardiovascular para apoyar los programas de prevención y promoción del modelo de salud. Modelo basado en datos clínicos de pacientes que hayan presentado alguna afección cardiovascular. El objetivo es apoyar de manera eficiente a los programas de prevención y promoción detectando de manera oportuna los pacientes con una mayor probabilidad de sufrir un episodio de afección cardiaca, y de esta manera actuar de forma anticipada preservando el estado de salud del paciente.

## 13. Referencias

- [cou-ml] <https://www.coursera.org/learn/machine-learning>
- [and-ml] <https://class.coursera.org/ml-005/lecture>
- [wik-ml] [https://es.wikipedia.org/wiki/Aprendizaje\\_autom%C3%A1tico](https://es.wikipedia.org/wiki/Aprendizaje_autom%C3%A1tico)
- [goo-news] <https://news.google.com/>
- [pred-game] <http://data-informed.com/predict-winners-big-games-machine-learning/>
- [spe-rec] <http://research.google.com/pubs/SpeechProcessing.html>
- [fra-dec] <https://www.research.ibm.com/foiling-financial-fraud.shtml>
- [pri-ml] [http://www.cs.princeton.edu/courses/archive/spr08/cos511/scribe\\_notes/0204.pdf](http://www.cs.princeton.edu/courses/archive/spr08/cos511/scribe_notes/0204.pdf)
- [nir-wso2-ml] <http://nirmalfdo.blogspot.com.co/2015/07/sneak-peek-into-wso2-machine-learner-10.html>
- [wso2-prod] <http://wso2.com/products/machine-learner/>



**SUNQU**

[www.sunqu.net](http://www.sunqu.net)