

Proyecto 1 Inteligencia Artificial

Juan Pablo Estavillo Urrea

Juan Mario Ochoa Navarro

Juan López Pelayo

Métodos de Inteligencia Artificial

ITESO A.C

Índice / Tabla Contenido

[Objetivo](#_Toc22593503)

[Marco teórico](#_Toc22593504)

[Algoritmo de Clasificación Supervisado: K Vecinos más cercanos](#_Toc22593505)

[Algoritmo de Clasificación No Supervisado: K medias](#_Toc22593506)

[Desarrollo](#_Toc22593507)

[**Primera forma normal**](#_Toc22593508)

[**Segunda forma normal**](#_Toc22593509)

[**e)** Proporcionar los tres modelos de clasificación y sus medidas de desempeño.](#_Toc22593510)

[Conclusiones](#_Toc22593511)

[Bibliografía](#_Toc22593512)

# 

# Objetivo

El objetivo del primer proyecto de inteligencia artificial es el de utilizar los datos proporcionados por el profesor, descubrir qué variables son realmente las que tienen peso en el problema, limpiar los datos, clasificarlos y realizar una predicción sobre que empresas con ciertas características se irían a default o no.

Este proyecto consiste en presentar evaluar empresas y predecir si se irán a default de pago o no con el apoyo de redes neuronales estáticas. La intención es que, utilizando este tipo de algoritmos, logremos encontrar un modelo de buena calidad que prediga con los datos proporcionados. Los modelos se realizarán en matlab.

# Marco teórico

Las redes neuronales estáticas son modelos computacionales que consisten en unidades llamadas neuronas por las cuales pasan los datos sometiéndose a varias operaciones matemáticas con el fin de dar un resultado de salida como respuesta para el problema en cuestión. Los datos pasan por cada capa (unida por un enlace) en la cual se multiplican por un peso el cual puede incrementar a disminuir.

Este método de redes neuronales aprende por sí mismo a través de iteraciones buscando disminuir una función de pérdida. En este caso en lugar de una función se busca clasificar con el fin de predecir con una mayor exactitud que va a ocurrir con las empresas en cuestión.

La clasificación no supervisada es un proceso mediante el cual se definen clases de objetos caracterizados por distintas variables (reales o discretas) sin conocer a qué clase pertenece cada objeto, es decir, la clasificación supervisada busca crear clases de equivalencia en donde los elementos que pertenezcan a la misma clase presenten ciertas características en común (homogeneidad); pero a su vez las distintas clases deben ser disjuntas (heterogeneidad).

La clasificación supervisada, también denominada reconocimiento de patrones, cada objeto está caracterizado por distintas variables (reales o discretas), así como la clase a la que pertenece cada objeto. Por lo tanto, el objetivo de la clasificación supervisada consiste en encontrar un modelo (formular una regla) para asignar un nuevo objeto a una de las categorías existentes en base a sus características. La muestra de objetos que sirve para construir el clasificador se le denomina conjunto de entrenamiento, ya que a partir de ella se determina la estructura y los parámetros del clasificador.

La clasificación supervisada juega un papel fundamental en las finanzas, ya que si se tienen distintos datos financieros de una empresa (variables) se puede determinar si ésta es candidata para sufrir una quiebra (entrar en default) o no.

Para realizar una clasificación adecuada no basta con sólo usar todos los datos con los que se dispone, es habitual en el ámbito de las finanzas dar un tratamiento previo a dichos datos y hacer una selección de las variables que más influyen en el fenómeno a caracterizar.

Los sistemas de aprendizaje supervisados son aquellos en los que, a partir de un conjunto de ejemplos de los que conocemos su valor objetivo (conjunto de entrenamiento), ya sea como clasificación o como regresión, intentamos encontrar una función que permita asignar un valor objetivo a ejemplos que el sistema no ha visto anteriormente (y de los que, generalmente, no conocemos el valor de salida correcto). En contra, los sistemas de aprendizaje no supervisados son aquellos en los que no disponemos de una salida esperada que asociar a la batería de ejemplos con la que trabajamos, sino que únicamente a partir de las propiedades de los ejemplos intentamos dar una agrupación o caracterización (clasificación, clustering) de los ejemplos según la similitud entre sus propiedades.

Aunque estas definiciones son generales y contienen entre las dos la práctica totalidad de los algoritmos de aprendizaje más comunes, en lo que sigue nos vamos a centrar en dos algoritmos de clasificación concretos, uno de cada tipo, que pueden servir de representación para entender su comportamiento más general, y que nos permitirán también introducir los conceptos generales de matriz de confusión, error, sobre-aprendizaje, etc.

# 

# Algoritmo de Clasificación Supervisado: K Vecinos más cercanos

El algoritmo de los **k vecinos más cercanos** (**k-NN Nearest Neighbour**) es un sistema de clasificación supervisado basado en criterios de vecindad. En particular, k-NN se basa en la idea de que los nuevos ejemplos serán clasificados a la clase a la cual pertenezca la mayor cantidad de vecinos más cercanos del conjunto de entrenamiento más cercano a él.

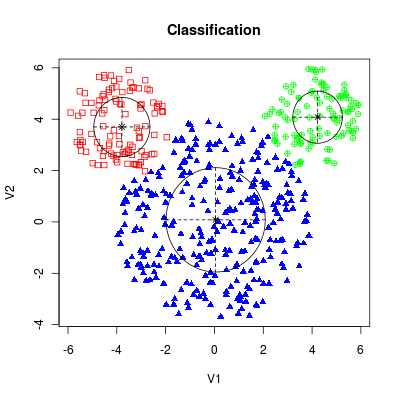
El **algoritmo del vecino más cercano** (aquel que asigna a una nueva muestra la clasificación de las muestras de ejemplo más cercana) explora todo el conocimiento almacenado en el conjunto de entrenamiento para determinar cuál será la clase a la que pertenece una nueva muestra, pero únicamente tiene en cuenta el vecino más próximo a ella, por lo que es lógico pensar que es posible que no se esté aprovechando de forma eficiente toda la información que se podría extraer del conjunto de entrenamiento.

Con el objetivo de resolver esta posible deficiencia surge la regla de los k vecinos más cercanos (k-NN), que es una extensión en la que se utiliza la información suministrada por los k ejemplos del conjunto de entrenamiento más cercanos.

## 

# 

# Algoritmo de Clasificación No Supervisado: K medias

****

Como ejemplo de algoritmo de clasificación no supervisado vamos a explicar con brevedad el que, posiblemente, sea el más sencillo y extendido de todos ellos, el **algoritmo de las K-medias**, que es aplicable en los casos en que tengamos una inmersión de nuestros ejemplos en un espacio métrico.

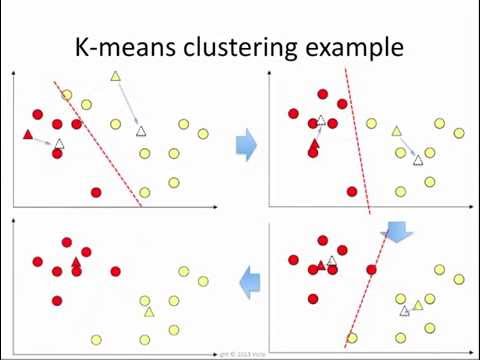
El algoritmo intenta encontrar una partición de las muestras en *K* K agrupaciones, de forma que cada ejemplo pertenezca a una de ellas, concretamente a aquella cuyo centroide esté más cerca. El mejor valor de *K* K para que la clasificación separe lo mejor posible los ejemplos no se conoce a priori, y depende completamente de los datos con los que trabajemos. Este algoritmo intenta minimizar la varianza total del sistema, es decir, si*ci*ci

es el centroide de la agrupación *i*i-ésima, y

{*xij*}{xji}

es el conjunto de ejemplos clasificados en esa agrupación, entonces intentamos minimizar la función:

∑*i*∑*jd*(*xij*,*ci*)2 ∑i∑jd(xji,ci)2

El algoritmo que se sigue es el siguiente:

1. Seleccionar al azar
2. *K*
3. K
4. puntos como centros de los grupos.
5. Asignar los ejemplos al centro más cercano.
6. Calcular el centroide de los ejemplos asociados a cada grupo.
7. Repetir desde el paso 2 hasta que no haya reasignación de centros (o su último desplazamiento esté por debajo de un umbral).

El algoritmo anterior es relativamente eficiente, y normalmente se requieren pocos pasos para que el proceso se estabilice. Pero en contra, es necesario determinar el número de agrupaciones a priori, y el sistema es sensible a la posición inicial de los

centros, haciendo que no consigan un mínimo global, sino que se sitúe en un mínimo local (algo muy común cuando se trabaja con un problema de optimización no convexo). Por desgracia, no existe un método teórico global que permita encontrar el valor óptimo de grupos iniciales ni las posiciones en las que debemos situar los centros, por lo que se suele hacer una aproximación experimental repitiendo el algoritmo con diversos valores y posiciones de centros. En general, un valor elevado de K

K hace que el error disminuya, pero a cambio se tiene un sobre entrenamiento que disminuye la cantidad de información que la agrupación resultante da

# Desarrollo

**a)**  Justificar por qué se decidió la selección de variables realizada. Si se siguió algún procedimiento formal describirlo.

## modelo 1

graficamos todas las variables con el eje de las y siendo el default, ordenando cada variable para buscar si existía una distribución de probabilidad que lograra que esta variable fuera influyente para el modelo, si no existía un sesgo en la distribución de probabilidad se omitía la variable, en caso contrario en la que existiera un sesgo se utiliza en el modelo para la predicción. las variables para utilizar en este modelo son date, share prices, currency exchange, hourly earnings, international trade in goods, cetes.

## modelo 2

Escogimos las 5 variables que consideramos cualitativamente más importante, basándonos en las variables que tienen una correlación (absoluta) mayor con respecto a default, esto con el fin de encontrar las 5 variables que más afectan el default sin sobresaturar de datos el modelo. Las variables fueron: date, retail trade, index of industry, international goods - exports y cetes

## modelo 3

Buscamos un modelo que considere todas las variables, ya que la correlación de los datos es baja no hemos descartado que todas las variables influyen y puede que un modelo robusto que utilice toda la base de datos tenga mejores resultados.

# b) Explicar qué es la normalización y para qué sirve.

La normalización es el proceso de organizar los datos de una base de datos. Se incluye la creación de tablas y el establecimiento de relaciones entre ellas según reglas diseñadas tanto para proteger los datos como para hacer que la base de datos sea más flexible al eliminar la redundancia y las dependencias incoherentes.

Los datos redundantes desperdician el espacio de disco y crean problemas de mantenimiento. Si hay que cambiar datos que existen en más de un lugar, se deben cambiar de la misma forma exactamente en todas sus ubicaciones. Un cambio en la dirección de un cliente es mucho más fácil de implementar si los datos sólo se almacenan en la tabla Clientes y no en algún otro lugar de la base de datos.

Normalización de Base de Datos, es el proceso de organizar los datos en una base de datos que incluye la creación de tablas y el establecimiento de relaciones entre ellas.

Este proceso es utilizado para ayudar a eliminar los datos redundantes.

Cinco formas de normalización (FN: Forma normal)

1FN: Eliminar grupos repetitivos

2FN: Eliminar datos redundantes

3FN: Eliminar columnas no depende de clave

4FN: Aislar Relaciones Múltiples Independientes

5FN: Aislar relaciones semánticamente relacionadas múltiples

**c) Si los datos contenidos en alguna variable se sometieron a una normalización indicar bajo qué método.**

## Primera forma normal

* Elimine los grupos repetidos de las tablas individuales.
* Cree una tabla independiente para cada conjunto de datos relacionados.
* Identifique cada conjunto de datos relacionados con una clave principal.

No use varios campos en una sola tabla para almacenar datos similares. Por ejemplo, para realizar el seguimiento de un elemento del inventario que proviene de dos orígenes posibles, un registro del inventario puede contener campos para el Código de proveedor 1 y para el Código de proveedor 2.

¿Qué ocurre cuando se agrega un tercer proveedor? Agregar un campo no es la respuesta, requiere modificaciones en las tablas y el programa, y no admite fácilmente un número variable de proveedores. En su lugar, coloque toda la información de los proveedores en una tabla independiente denominada Proveedores y después vincule el inventario a los proveedores con el número de elemento como clave, o los proveedores al inventario con el código de proveedor como clave.

## Segunda forma normal

* Cree tablas independientes para conjuntos de valores que se apliquen a varios registros.
* Relacione estas tablas con una clave externa.

Los registros no deben depender de nada que no sea una clave principal de una tabla, una clave compuesta si es necesario. Por ejemplo, considere la dirección de un cliente en un sistema de contabilidad. La dirección se necesita en la tabla Clientes, pero también en las tablas Pedidos, Envíos, Facturas, Cuentas por cobrar y Colecciones. En lugar de almacenar la dirección de un cliente como una entrada independiente en cada una de estas tablas, almacénese en un lugar, ya sea en la tabla Clientes o en una tabla Direcciones independiente.

### **La normalización** de una base de datos persigue varios objetivos, *principalmente reducir la redundancia de datos y simplificar las dependencias entre columnas, aplicándose de manera acumulativa*. Lo anterior quiere decir que la segunda forma normal incluye a la primera, la tercera a la segunda y así sucesivamente. Una base de datos que esté en segunda forma normal, por tanto, cumplirá las dos primeras reglas de normalización.

**d)** Justificar el cómo se eligió el grado del clasificador.

Utilizamos una gráfica de codo para el grado del clasificador, con el cual usamos un grado dos.

# **e)** Proporcionar los tres modelos de clasificación y sus medidas de desempeño.

modelo1

|  |  |
| --- | --- |
| accu | .9097 |
| pre | .6667 |
| rec | .25 |

modelo 2

|  |  |
| --- | --- |
| accu | .9129 |
| pre | 1 |
| rec | .1563 |

modelo 3

|  |  |
| --- | --- |
| accu | .9645 |
| pre | .8621 |
| rec | .7813 |

# Conclusiones

Ahora con este proyecto podemos apreciar como los datos obtenidos después de ser normalizados para poderlos manejar nos permite analizarlos de una manera más comparativa ya que si no fuesen normalizados serían cantidades diferentes para analizar, nuestro comportamiento en cuanto a la base de datos que obtuvimos tiene un comportamiento normal. La correlación de todos los datos nos hace darnos cuenta la relación que conllevan los datos entre sí, a pesar de que los datos default fueron incluidos también demuestra que hubo correlación en cuanto a cada una de las columnas, unas más que otras, si nos pusiéramos a hacer el modelo con las diferentes columnas y tomando en cuenta diferentes correlaciones entre los datos obtenidos, se hubiese encontrado correlaciones más cercanas, dándonos modelos más simples pero por lo mismo de que los datos eran distintivos en cuanto a sus cantidades, precios etc, cada uno de los modelos nos aparecen con diferente grado de correlación y pues por lo mismo el recall y la exactitud puede perder fuerza por lo mismo.

# Bibliografía

**Proyecto2.Práctica redes neuronales estáticas. Saúl Nuño. Departamento de Matemáticas y Física. Instituto Tecnológico y de Estudios Superiores de Occidente.**

**Clasificación Supervisada y no Supervisada (2018). Fernando Sancho Caparrini. Recuperado de:** [**http://www.cs.us.es/~fsancho/?e=77**](http://www.cs.us.es/~fsancho/?e=77)

**Fundamentos de la normalización de la base de datos. Recuperado de :** [**https://support.microsoft.com/es-mx/help/283878/description-of-the-database-normalization-basics**](https://support.microsoft.com/es-mx/help/283878/description-of-the-database-normalization-basics)

**Normalización de bases de datos. Recuperado de :** [**https://programas.cuaed.unam.mx/repositorio/moodle/pluginfile.php/872/mod\_resource/content/1/contenido/index.html**](https://programas.cuaed.unam.mx/repositorio/moodle/pluginfile.php/872/mod_resource/content/1/contenido/index.html)

**Normalización de bases de datos Recuperdado de:** [**http://www.marcossarmiento.com/2017/06/28/normalizacion-de-base-de-datos/**](http://www.marcossarmiento.com/2017/06/28/normalizacion-de-base-de-datos/)