*Universidad de La Laguna*

*Escuela de Doctorado y Estudios de Posgrado*

*Departamento de Ingeniería Informática y de Sistemas*

# Extracción de conocimiento en bases de datos

PRÁCTICA: CLASIFICACIÓN

*J. Marcos Moreno Vega*

UNIVERSIDAD DE LA LAGUNA

TUFTE-LATEX.GOOGLECODE.COM

---

**Objetivo:**
Construir y evaluar modelos de clasificación.

**Base de datos:**
homeLoanAproval.csv

**Lenguaje de programación:**
R o Python.

---

## *Concesión de hipotecas*

El mercado hipotecario mueve anualmente una cifra de negocios que alcanza los miles de millones de euros. Las entidades bancarias deben gestionar miles de solicitudes anuales y decidir qué hipotecas conceder y cuáles denegar. Conceder una hipoteca que luego no será devuelta supone un perjuicio económico para la entidad. Del mismo modo, denegar una hipoteca a un solicitante con capacidad para afrontar los pagos es perder una oportunidad de negocio.

Se desea construir un sistema basado en Analítica de datos que ayude en la tarea de concesión de los créditos. Para ello, se dispone de información detallada de los créditos gestionados en el pasado, así como de la decisión que se adoptó con cada solicitud.

## *Base de datos*

La base de datos homeLoanAproval.csv almacena información sobre 614 solicitudes de crédito hipotecario. El fichero consta de 12 variables, 11 de las cuáles son predictivas. En la tabla 1.1 se enumeran y describen todas las variables.

## *Tareas*

a) Preparar los datos. Desarrollar las tareas de prepararación de los datos necesarias para disponer de una base de datos apropiada para la tarea de clasificación. Entre las labores a realizar se encuentran las de visualización y comprensión de las variables, cambio de tipos, transformación e identificación de outliers y datos faltantes. También debe quedar claro si nos enfrentamos a una clasificación con clases balanceadas o no.

b) Analizar los datos. Construir un clasificador $k$-NN, un árbol de clasificación y un clasificador naive Bayes. Se propone considerar varios escenarios. En el primero, independientemente de que existan o no datos faltantes o outliers, se aplicará el correspondiente algoritmo de inducción sin tener en cuenta estos hechos. En el segundo escenario, se aplicarán técnicas para tratar con los outliers y los datos faltantes y se construirán los anteriores clasificadores. Debe evaluarse también el uso de técnicas para tratar con clases

| Variable | Descripción | Valores |
|---|---|---|
| LoanID | identificador de la solicitud | Alfanumérico |
| Gender | género del solicitante | Male/Female |
| Married | variable que indica si el solicitante es una persona casada | Yes/No |
| Dependents | número de personas a cargo del solicitante | Numérico |
| Education | nivel educativo del solicitante | Graduate/Not Graduate |
| SelfEmployed | variable que indica si el solicitante es autónomo o no | Yes/No |
| ApplicationIncome | ingresos del solicitante | Numérico |
| CoapplicationIncome | ingresos del segundo solicitante del crédito | Numérico |
| LoanAmount | importe del crédito hipotecario | Numérico |
| LoanAmountTerm | periodo del crédito (en meses) | Numérico |
| PropertyArea | lugar en que se ubica la propiedad hipotecada | Rural/Urban/Semiurban |
| LoanStatus | variable que indica si el crédito fue aprobado | Yes/No |

Tabla 1.1: Base de datos homeLoanAproval.csv

desbalanceadas y el efecto que tienen estas técnicas en la calidad de los clasificadores obtenidos.

### *Qué debe presentar el alumnado*

a) Cuaderno Jupyter, debidamente comentado, usado para preparar y analizar los datos.

b) Una memoria en formato pdf en la que se describan brevemente el análisis realizado y el modelo de clasificación obtenido para ayudar a los responsables de las entidades bancarias.

c) La memoria debe incluir también gráficas y tablas que muestren la calidad de los clasificadores obtenidos y que permitan identificar cuál es el más apropiado.