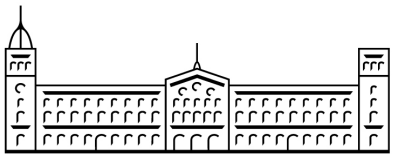

Treball final de grau

GRAU DE MATEMÀTIQUES

Facultat de Matemàtiques
Universitat de Barcelona

CREACIÓ D'UN PILOT COMPARATIU DE TÈCNIQUES DE DETECCIÓ DE FRAU FISCAL

Alba Conde Rodríguez

Director:	Dr. Josep Fortiana Gregori
Realitzat a:	Departament Probabilitat, Lògica i Estadística. UB

Barcelona, 30 de juny de 2015

Abstract

The present project consists of the study and analysis of several Statistical Classification methods oriented towards their application to real estate dealing contracts data with the aim of detecting cases of fiscal fraud where the declared transaction value is below a statutory price level.

Resum

Aquest projecte consisteix en l'estudi i anàlisi d'un conjunt de mètodes estadístics de classificació orientat a l'aplicació en dades provinents de contractes de compravenda del sector immobiliari, amb l'objectiu de detectar casos de frau fiscal on el valor de la transacció declarada estigui per sota del nivell de preus legal.

Agraïments

Abans de començar qualsevol agraïment, aquest treball no hagués estat possible sense l'existència de totes les persones que s'han trobat al meu voltant durant l'execució d'aquest. Per això, els hi dono les gràcies per aportar-me els ànims quan els he necessitat i per intentar-me ajudar quan han vist que no trobava el camí. Per ser una persona que no només ho ha intentat, sinò que m'ha ajudat sempre que li he demanat, dono les gràcies al meu tutor Josep Fortiana per la seva paciència i la seva dedicació.

Índex

1 Introducció

Què és el frau fiscal?

El frau fiscal cada cop ha anat agafant més protagonisme ja que ha ajudat a agreujar la situació econòmica mundial, especialment als països europeus. Aquest fenomen es troba fortament enllaçat amb l'economia submergida, comparteixen la mateixa base: la infradeclaració de les bases, així com també el debilitament d'ingressos de l'Estat. Essent més acurats, es pot dir que es tracta d'una modalitat d'evasió fiscal premeditada, en general punible pel dret penal que inclou la presentació de documents o de declaracions falses.

Repercussió

"Cada dia es perd al voltant d'una cinquena part dels diners públics en Europa degut al frau i a l'evasió fiscal"[6]

La quantitat de diners que escapen del control de fisc degut al frau i l'evasió fiscal en Espanya són superior als 250.000 milions d'euros per any. Segons Gestha[1],per conceptes tributaris, d'aquests milions s'estimen que uns 90.000 anuals provenen de l'evasió d'impostos i cotitzacions socials i uns 17.176 anuals són pèrdues per defraudació en l'IVA[2], un 1.6% del PIB[3] nacional. Des de la crisi en el 2008 fins al 2013, la taxa ha augmentat del 17,85% fins el 25% del PIB, situant Espanya a la capçalera del frau entre les grans potències europees - per davant d'Itàlia. Aquest increment no solament és degut en gran part a la forta sensació de corrupció que viu la població espanyola, sinó també a la poca implicació en l'erradicació del frau per part del Govern.

A nivell autonòmic es detecta diferenciació en els nivells de corrupció, representats en gràfics a [7], així com també segons la font del frau fiscal. Per aquesta raó, també hi ha diferents mesures de prevenció i correcció segons l'origen d'aquest frau, que van des de la promoció de la participació ciutadana i creació d'una ética ciutadana fins el tractament del frau tributari. En aquest últim cas, es poden dividir en activitats empresarials, lloguers, compra-venda d'immobles, empreses, vehicles.

Objectius

En aquest treball apliquem tècniques d'aprenentatge automàtic (*machine learning*) a la detecció del frau fiscal en la compra-venda d'immobles consistent a declara com a preu de la transacció un import inferior al valor real del bé.

[1]Sindicat de tècnics del Ministeri d'Hisenda
[2]Impost sobre el Valor Afegit
[3]Producte Interior Brut

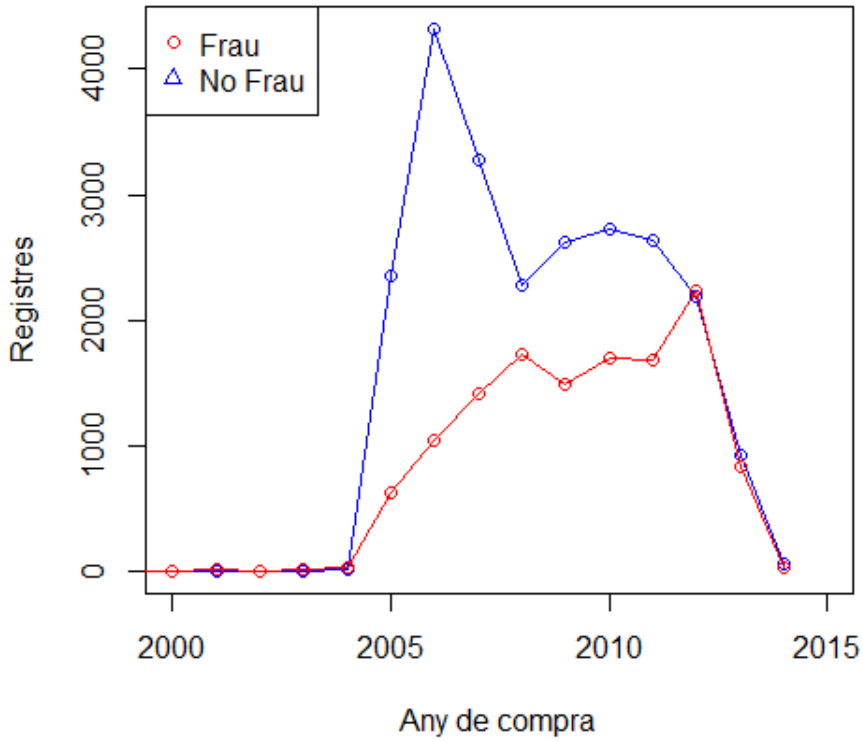

Figura 1: Relació entre fraudulents - no fraudulents

La figura 1, ens dóna una idea de com es distribueixen les compra-venda fraudulentes respecte les que no ho són, partint de les dades inicials d'aquest treball.

Confidencialitat

El contingut d'aquest TFG és de caràcter aplicat i vinculat a unes pràctiques externes realitzades a l'empresa INDRA SISTEMAS ubicada c.de Roc Boronat, 133, Barcelona, amb el codi postal 0818.
Per aquesta raó, en virtut del conveni de confidencialitat [10] signat amb l'empresa abans esmentada, les dades amb les quals s'han calculat els resultats numèrics que hi apareixen han estat sotmeses a procediments d'anonimització. (Conveni de práctiques núm. 81/2014-15 punt 10.Acord de confidencialitat)

2 Conceptes previs

Amb l'objectiu d'agilitzar la lectura dels conceptes que defineixen el gruix dels models, a continuació s'exposaran les definicions bàsiques relacionades amb el tema a tractar.

2.1 De teoria de probabilitats

Teorema (Teorema de Bayes)**.** *Sigui* $\{A_1, A_2, \ldots, A_i, \ldots, A_n\}$ *un conjunt de esdeveniments, d'un espai de probabilitat* $(\Omega, \mathcal{F}, \mathcal{P})$*, mútuament excloents,* $A_i \cap A_j = \emptyset$ *per a* $i \neq j$*, i exhaustius,* $\bigcup_{i=1}^{n} A_i = \Omega$*, i tals que la probabilitat de cadascun d'ells és diferent de zero. Sigui* B *un esdeveniment qualsevol del que es coneixen les probabilitats condicionals* $P(B|A_i)$*. Llavors, la probabilitat* $P(A_i|B)$ *ve donada per l'expressió:*

$$P(A_i|B) = \frac{P(B|A_i)P(A_i)}{P(B)} = \frac{P(B|A_i)P(A_i)}{\sum_{i=1}^{n} P(B|A_i)P(A_i)} \propto P(B|A_i)P(A_i)$$

on:

- $P(A_i)$ *són les probabilitats inicials o* a priori.
- $P(B|A_i)$ *és la probabilitat de* B *en la hipòtesis* A_i*.*
- $P(A_i|B)$ *són les probabilitats finals o* a posteriori.

Definició 1. *Sigui* $X = \{X_1, X_2\}$ *i* Y*, un conjunt de variables aleatòries definides en un mateix espai de probabilitat. Diem que* X_1 *i* X_2 *són* ***condicionalment independents*** *donada* Y*, si, i només si, els esdeveniments* $[X_1 = x_1]$ *i* $[X_2 = x_2]$ *són condicionalment independents, donat* $[Y = y]$ *per a qualsevol parella de valors,* x_1, x_2 *i per qualsevol* y*, complint* $P(y) > 0$*, és a dir,*

$$\forall i, \forall j, \forall k,\ P(X_1 = x_i, X_2 = x_j|Y = y_k) = P(X_1 = x_i|Y = y_k) \cdot P(X_2 = x_j|Y = y_k)$$

Definició 2. *Sigui* $X = \{X_1, X_2, \ldots, X_n\}$ *una mostra aleatòria simple de una v.a. amb densitat* $f(x|\theta)$ *en que* $\theta \in \Theta$*, l'espai de paràmetres.*

Anomenem ***funció de versemblança*** *la densitat conjunta del vector dels valors mostrals.*

$$\forall x = (x_1, \ldots, x_n) \text{ en la mostra, es denota } f_n(x|\theta) = f_n(x_1, x_2, \ldots, x_n|\theta)$$

En particular, si $X_1, X_2, \ldots, X_n$ són i.i.d,

$$\mathcal{L}(x_1, \ldots, x_n; \theta) = f_n(x_1, x_2, \ldots, x_n; \theta) = \prod_{i=1}^{n} f(x_i; \theta)$$

A la pràctica, generalment, s'aplica el logaritme a $\mathcal{L}$,

$$\ell(x;\theta) \equiv \ln \mathcal{L}(x_1, x_2, \ldots, x_n; \theta) = \sum_{i=1}^{n} \ln f(x_i;\theta)$$

Definició 3. *El mètode de màxima versemblança busca el valor θ que maximitza $\ell(\theta; x)$. Aquest mètode defineix l'**estimador de Màxima Versemblança (MLE)***

$$\hat{\theta}_{MLE} = \arg\max_{\theta\in\Theta} \ell(x_1, \ldots, x_n; \theta)$$

Ara, suposem que la distribució *a priori* g de θ existeix. Això ens permet tractar θ com una variable aleatòria, com a la teoria Bayesiana. Aleshores la distribució posterior de θ és:

$$\frac{f(x|\theta)g(\theta)}{\int_{v\in\Theta} f(x|v)g(v)dv} \tag{2.1}$$

on g és la funció de densitat de θ, Θ és el domini de g.

Definició 4. *Definirem com a **estimació Màxima a Posteriori (MAP)** l'estimador $\hat{\theta}_{MAP}$ que maximitza la distribució* posterior (2.1).

2.2 D'aprenentatge automàtic

Definició 5. *L'aprenentatge supervisat consisteix en, donat $(x_1, y_1), \ldots, (x_n, y_n)$ amb $x_i \in X \subseteq \mathbb{R}^p$ i $y_i \in Y \subseteq \mathbb{R}^k$ $\forall i$, el conjunt d'entrenament, buscar una funció M tal que, la funció, algorisme o mètode de predicció,*

$$M : X \to Y \text{ compleixi } M(x_i) \approx y_i \text{ (aproximadament)}$$

on $M(x_i) \approx y_i$ vol dir que, la funció de predicció M ha d'obtenir prediccions $\hat{y}_i$, aproximades a les observacions y_i, de manera que quan s'apliqui a noves observacions x produeixi bones aproximacions als corresponents valors y, no observats.

Segons quin tipus de variable sigui Y l'anomenarem d'una forma o d'una altra:

- *Si Y és de tipus quantitatiu parlem de **regressió**.*
- *Si $Y \in \{C_1, C_2, \ldots, C_k\}$, és a dir, té valor qualitatius, parlem de **classificació**.*

2.3 Del paradigma Bayesià en estadística

Tot i que en aquest treball no fem ús exhaustiu de la perspectiva Bayesiana, exceptuant algun apartat en la part d' aproximació teòrica com el Naive Bayes, fem servir alguns termes tècnics que aclarirem a continuació.

Si, en comptes del paradigma de l'Estadística clàssica (Neyman - Pearson), adoptem el paradigma Bayesià, els paràmetres d'un model estadístic passen a ser variables aleatòries, amb les seves distribucions de probabilitat. En particular, la distribució *a priori* (abans d'incorporar la informació continguda a les observacions) i la distribució *a posteriori* (després). Aquestes dues distribucions, es relacionen amb la versemblança de les observacions mitjançant la fórmula de Bayes (Definició 2.1).

3 Aproximació teòrica

En aquest apartat s'explicaran els conceptes necessaris per a descriure els models que s'apliquen en aquest treball.

Cadascuna d'aquestes tècniques de classificació pertany al grup de models d'aprenentatge supervisat (Definició 5), és a dir, necessiten d'un conjunt de dades d'entrenament per a poder generar el model i posteriorment, utilitzar-lo en un altre conjunt de dades per poder estimar la eficàcia del mètode. Malgrat això, cadascun d'ells té diferent naturalesa:

- *Naive Bayes*: és un model que té com a base el càlcul de probabilitats condicionades, la teoria Bayesiana.
- *Bagging*: és una técnica amb la que es combinen diferents classificadors, en aquest cas arbres CART. Per tant té com a base teoria de grafs.
- *Màquines de vectors suport*: és un mètode que té una base geomètrica.

3.1 Naive Bayes

Naive Bayes és un algorisme de classificació que es basa en el Teorema de Bayes, (Teorema 2.1). La *naïvété* (ingenuïtat) del mètode és presuposar la independència condicional entre variables predictores, $X_1, \ldots, X_n$, donada Y. Naturalment, aquesta hipòtesi és molt forta, i generalment no es compleix a la pràctica. És, però, una primera aproximació que pot ser útil. Aquesta característica és la raó principal del seu èxit, la simplificació de la representació de $P(X|Y)$, i el problema d'estimació d'aquest conjunt partint d'un conjunt de dades d'entrenament.

Suposant la independència condicional (Definició 1), tenim

$$P(X_1, \ldots, X_n|Y) = \prod_{i=1}^{n} P(X_i|Y), \tag{3.1}$$

les probabilitats que calcularem partint del conjunt de dades inicial.

Tornant al classificador Naive Bayes, el nostre objectiu és entrenar-lo de manera que ens retorni la distribució de probabilitat de cada possible valor de Y per cadascuna de les X. La fórmula probabilística que ens retornarà el possible valor de Y a la posició i-èsima és

$$P(Y = y_i|X_1, \ldots, X_n) = \frac{P(X_1, \ldots, X_n|Y = y_i)P(Y = y_i)}{\sum_j P(X_1, \ldots, X_n|Y = y_j)P(Y = y_j)} \tag{3.2}$$

on el sumatori pren tots els possibles valors de Y.

Aplicant l'assumpció d'independència a les X_k donades per Y, obtenim a partir de (3.2) la següent equació:

$$P(Y = y_i|X_1, \ldots, X_n) = \frac{\prod_k P(X_k|Y = y_i)P(Y = y_i)}{\sum_j P(Y = y_j)\prod_k P(X_k|Y = y_j)} \tag{3.3}$$

Amb la fórmula (3.3), a partir d'un nou conjunt de dades $X^* = \{X_1^*, \ldots, X_n^*\}$ i amb les distribucions P(Y) i $P(X_i|Y)$ calculades del conjunt de dades d'aprenentatge, retornarà la probabilitat de Y que té en cadascun dels possibles valors. Com només estem interessats en el valor superior:

$$Y \leftarrow \arg\max_{y_i} \frac{\prod_k P(X_k|Y = y_i)P(Y = y_i)}{\sum_j P(Y = y_j)\prod_k P(X_k|Y = y_j)}$$

que es pot simplificar, ja que el denominador no depèn de y_i.

$$Y \leftarrow \arg\max_{y_i} P(Y = y_i)\prod_k P(X_k|Y = y_i)$$

A continuació anem a analitzar el funcionament del classificador Naive Bayes segons si el conjunt de valors d'entrada és categòric o numèric.

3.1.1 Generació de probabilitats de dades d'entrada de tipus discret

Suposem que les n variables X_k prenen K possibles valors (classes) amb Y pertanyent a un conjunt de g elements (classes).

Veurem com s'estima el valor de

$$\tau_{kji} = P(X_k = x_{kj}|Y = y_i),$$

per a cada X_k amb valor x_{kj}, i a la vegada, per a cada possible valor y_i de Y, complint que la suma d'aquestes probabilitats per a cada parell de k, i valors és 1, $\sum_j \tau_{kji} = 1$.

Per altra banda, també hem d'estimar la probabilitat *a priori* d'Y:

$$\pi_i = P(Y = y_i)$$

Cadascun d'aquests valors els podem estimar usant l'estimador de màxima versemblança (MLE), (Definició 3), basat en calcular les freqüències relatives de diferents esdeveniments, o fent ús de l'estimador Bayesià *Màxim-a-posteriori* (MAP), (Definició 4).

L'estimació de màxima versemblança de τ_{kji} que calculem a partir del conjunt de dades d'entrada és

$$\hat{\tau}_{kji} = (X_k = x_{kj}|Y = y_i) = \frac{\#(X_k = x_{kj} \cap Y = y_i)}{\#(Y = y_i)}$$

El risc que prenem utilitzant aquest estimador és obtenir $\hat{\tau}_{kji} = 0$, en cas d'inexistència d'alguns valors en el conjunt d'aprenentatge. Per evitar aquestes situacions, s'utilitza un altre estimador abans esmentat, l'estimador MAP.

Aquest estimador s'encarrega de fer que l'optimització es "suavitzi", és a dir, l'expressió de $\hat{\tau}_{kji}$ seria de la forma:

$$\hat{\tau}_{kji} = (X_k = x_{kj}|Y = y_i) = \frac{\#(X_k = x_{kj} \cap Y = y_i) + m}{\#(Y = y_i) + mK}$$

on m determina el nivell de suavitat que li apliquem. L'expressió anterior, pertany a l'estimador MAP per a τ_{kji} assumint la distribució *prior* de Dirichtlet sobre els paràmetres de τ_{kji}, amb paràmetres igualment distribuïts. Si $m = 1$, s'anomena Laplace *smoothing*.

Ara, estimem el valor de π_i, que aplicant MLE tenim,

$$\hat{\pi}_i = (Y = y_i) = \frac{\#(Y = y_i)}{\#(Y)}$$

Com anteriorment, existeix la opció de calcular l'estimació MAP basada en la distribució *a priori* de Dirichtlet sobre els paràmetres de π_i, en comptes de la MLE, obtenint

$$\hat{\pi}_i = (Y = y_i) = \frac{\#(Y = y_i) + m}{\#(Y) + mg}$$

amb g els nombre de valors diferents que pot prendre Y, i amb m com a determinant del nivell de suavitat usant distribucions *a priori*.

3.1.2 Generació de probabilitats de dades d'entrada de tipus continu

En el cas continu, tot i que necessitarem representar les distribucions d'una manera diferent a la que hem vist, podrem fer ús d'algunes de les equacions ja vistes com (3.1) i (3.3), ja que són les fonamentals del classificador Naive Bayes.

Una aproximació molt habitual és assignar la distribució Gaussiana a cada variable contínua X_k, $X_1, \ldots, X_k \sim N(\mu, \sigma^2)$ i.i.d, per a cada possible valor de Y, y_i.

En aquest cas, estimem la mitjana i la desviació estàndar de cadascun d'ells per a cada X_i i cada possible valor y_i:

$$\mu_{ki} = E[X_k|Y = y_i]$$
$$\sigma^2_{ki} = E[(X_k - \mu_{ki})^2|Y = y_i]$$

Com en l'apartat anterior, per a estimar el valor més probable de y_i, a més de calcular la mitjana i la desviació estàndar, hem d'estimar les probabilitats *a priori* de Y,

$$\pi_i = P(Y = y_i)$$

Per tant, estem assumint que les dades de $X = \{X_1, \ldots, X_n\}$ tenen distribució Gaussiana depenent del valor de la variable Y i, a més a més, per la condició Naive Bayes, X_i són condicionalment independents (Definició 1).

Un altre cop, podem escollir entre l'estimador MLE o MAP per a estimar els paràmetres μ_{ki} i σ^2_{ki}.

L'estimació per MLE de μ_{ki} és:

$$\hat{\mu}_{ki} = \frac{1}{\sum_p \delta(Y^p = y_i)} \sum_p X^p_k \delta(Y^p = y_i)$$

on p fa referència a l'observació p-èsima del conjunt d'entrenament, i

$$\delta(Y = y_i) = \begin{cases} 1 \text{ si } Y = y_k \\ 0, \text{ altrament.} \end{cases}$$

Apliquem MLE també a σ^2_{ki},

$$\hat{\sigma}^2_{ki} = \frac{1}{\sum_p \delta(Y^p = y_i)} \sum_p (X^p_k - \hat{\mu}_{ki})^2 \delta(Y^p = y_i)$$

Consultar [3] per a més informació.

3.2 Predicció amb arbres i *Bagging*

En aquest apartat, veurem com es poden fer prediccions amb arbres de decisió i, a més, com pot millorar aquesta predicció el mètode d'ensemble *bagging*.

Els conceptes d'arbre, node, node arrel, pare, fulles, graf dirigit, entre d'altres conceptes bàsics de Teoria de Grafs, es donen per coneguts.

3.2.1 Arbres CART

Els mètodes basats en arbres, primer separen l'espai en un conjunt de regions i després ajusten un model en cadascuna d'elles. És un concepte simple però potent.

En aquest apartat, descriurem el mètode més popular basat en arbres de classificació i regressió anomenat CART. Nosaltres ens centrarem en el cas de classificació.

L'objectiu que es persegueix durant el procés de construcció, és adquirir un arbre binari minimitzant l'error de predicció en cada fulla.

Construcció d'un arbre de classificació

Sigui $(x_1, y_1), \ldots (x_n, y_n)$ amb $x_i \in \mathbb{R}^p$, el nostre conjunt d'entrenament, amb y_i pertanyent a un conjunt de g elements (classes). L'algorisme necessita decidir automàticament les variables de separació, així com també, els valors per als quals separar.

Comencem la divisió, a partir de totes les dades d'entrenament, considerant una variable de separació j i un punt de separació s, i definint un parell de plans:

$$R_k(j,s) = \{x_i \in R_k | x_{ij} \leq s\} \text{ i } R_k^{'}(j,s) = \{x_i \in R_k | x_{ij} \geq s\},$$

que ens separaran les dades en dos conjunts.

Volem trobar la variable j i el valor s que resolguin

$$\min_{j,s}[N_{R_k(j,s)} \cdot E_{R_k(j,s)} + N_{R_k'(j,s)} \cdot E_{R_k'(j,s)}], \tag{3.4}$$

amb $N_{R_k} = \#\{i : x_i \in R_k\}$ i E_{R_k} la funció d'impuresa del node k que representa R_k, és a dir, que minimitzi la funció d'impuresa.

Primer, especifiquem com es fa el modelatge de cada regió R_k. Per a un node m, representant la regió R_m amb N_{R_m} observacions, tenim

$$\hat{p}_{mg} = \frac{1}{N_{R_m}} \sum_{x_i \in R_m} I(y_i = g),$$

la proporció d'observacions de clase g en el node m. Aleshores, classifiquem les observacions corresponents al node m com $g(m) = \arg\max_g \hat{p}_{mg}$, la classe majoritària al node m.

Respecte a la funció d'impuresa $E_{R_k(j,s)}$, hi han diferents mesures:

- Error de mala classificació : $\frac{1}{N_{R_m}} \sum_{i \in R_m} I(y_i \neq g(m)) = 1 - \hat{p}_{mg(m)}$
- Índex de Gini : $\sum_{g \neq g'} \hat{p}_{mg}\hat{p}_{mg'} = \sum_{g=1}^{G} \hat{p}_{mg}(1 - \hat{p}_{mg})$
- Entropia : $-\sum_{g=1}^{G} \hat{p}_{mg} \log \hat{p}_{mg}$

D'aquestes mesures anteriors, la utilitzada als arbres CART és l'índex de Gini.

Tornant a la condició (3.4), un cop hem trobat els valors j i s que la compleixen, els nodes representats per R_k i $R_{k'}$ passen a formar part de l'arbre de classificació, i de manera recursiva es repeteix el procés per a cadascun d'ells.

Arribats a aquest punt, un altre aspecte a tenir en compte és la dimensió de l'arbre, ja que segons si és massa gran o és massa petit afectarà a les prediccions per sobre ajustament o per falta d'informació respectivament. Per evitar que això passi, s'aplica un criteri d'aturada com la homogeneïtat de classes en el node o un

número mínim d'elements en el node; posteriorment s'ajusta l'arbre utilitzant el *cost-complexity pruning* com explicarem a continuació.

Definim $T \subset T_0$, un arbre fruit de la unió de nodes intermitjos de T_0. Indexem cada fulla amb m, representant respectivament la regió R_m. Sigui $|T|$, el nombre de nodes terminals de T, definim el criteri de complexitat del cost com

$$C_\alpha(T) = \sum_{m=1}^{|T|} N_{R_m} E_{R_m}(T) + \alpha|T|$$

Per a cada α, busquem l'arbre T_α més petit amb $C_\alpha(T)$ mínim. Consultar [1] per a més informació.

3.2.2 *Bootstrap*

El *bootstrap* és un mètode de computació intensiva que genera a partir d'unes observacions més conjunts de dades, mostres. Aquest sistema de remostratge té alguns avantatges com l'extracció de més informació útil sense haver de fer suposicions simplificadores sobre la distribució de probabilitat que segueixen les observacions.

Remostreig usant *bootstrap*

Sigui $Z = \{Z_1, \ldots, Z_n\}$ el conjunt de dades inicials, amb $Z_i = (x_i, y_i)$ amb $x_i \in \mathbb{R}^p$. Generem N remostres, del mateix tamany que Z.

Cadascuna d'aquestes remostres s'obtenen seleccionant elements del conjunt de dades inicials amb reemplaçament i probabilitat 1/n.

Llavors per cada conjunt generat, computem l'arbre CART associat, i d'aquesta forma, aplicant CART a cadascuna de les N remostres, ens retornarà la mostra *bootstrap* de l'arbre.

3.2.3 *Bagging*

Després d'introduir diferents conceptes relacionats amb aquest mètode d'ensemble, veurem en què consisteix el mètode d'ensemble *bagging*.

Bagging, o també anomenat ***B**ootstrap **Agg**regat**ing***, és un classificador que té per objectiu millorar les prediccions d'algorismes d'aprenentatge supervisat inestables, principalment reduint la variància. Els algorismes on s'aplica aquest classificador, són, en general, arbres de decisió (CART), tot i que es pot implementar en altres mètodes tant de regressió com de classificació.

3.3 Màquines de Vectors Suport

Definició 6. *SVM és un classificador que fa servir l'hiperplà amb marge màxim per a separar per classes les dades. Quan aquestes classes no són linealment separables*

en l'espai $\mathbb{X}$*, transforma les dades d'entrada a un espai de dimensió superior* $\mathbb{T}$ *conegut amb el nom d'*espai de configuracions *en el que sí podem trobar l'hiperplà separador.*

Per a introduir els conceptes, començarem amb el cas linealment separable. Aquest cas és resolt per LSVM (*Linear Support Vector Machines*).

Els passos generals que seguirem en la resolució del problema, en les diferents situacions, seran els següents:

- Descripció geomètrica del problema, on veurem els diferents elements que intervenen en el cas on estiguem treballant.
- Plantejament del problema d'optimització i adaptació de les condicions per trobar la solució.
- Recerca dels paràmetres òptims β i β_0 en dues etapes:
  - Solució del problema *primal* d'optimització : trobant el marge màxim minimitzant (3.8) fent ús de la funció Lagraniana.
  - Solució del problema dual respectiu, $Q(\alpha)$.

3.3.1 Descripció geomètrica de la LSVM

Sigui $(x_1, y_1), \ldots, (x_n, y_n)$ amb $x_i \in \mathbb{R}^p$, el nostre conjunt d'aprenentatge, amb les y_i pertanyent a un conjunt de g elements (classes). La teoria de LSVM es basa en l'hiperplà L amb representació

$$y = f(x) = \beta_0 + \beta^t x \tag{3.5}$$

on $\beta = (\beta_1, .., \beta_p)$ és el vector dels pesos, β_0 és el biaix i x és el vector amb els p predictors quantitatius.

L'hiperplà L (3.5), en $y = 0$, complirà les següents propietats:

1. Per a qualsevol parell de punts $x_1, x_2 \in$ L, $\beta^t(x_1 - x_2) = 0$. β és ortogonal a L, per tant $\beta^* = \frac{\beta}{\|\beta\|}$ és el vector normal a L.
2. La distància ortogonal de L a 0 és $\frac{-\beta_0}{\beta}$.
3. La distància ortogonal d'un punt x a L és $\frac{\beta^t \cdot x}{\|\beta\|}$.
4. Per qualsevol punt $x_0 \in L$, $\beta^t x_0 = -\beta_0$.
5. La distància des de qualsevol punt x a L és

$$\beta^{*t}(x - x_0) = \frac{1}{\|\beta\|}(\beta^t x + \beta_0) = \frac{1}{\|f'(x)\|} f(x) \tag{3.6}$$

Per tant, $f(x)$ és proporcional a la distància de x a $f(x) = 0$.

Consultar [1] per a més informació.

Exemple 1. Suposem $g = 2$ i $y \in \{-1, 1\}$ les possibles classes, definirem l'hiperplà L, separador de les classes, com:

$$\beta^t \cdot x_i + \beta_0 \geq 0, \quad \text{si } y_i = +1$$

$$\beta^t \cdot x_i + \beta_0 \leq 0, \quad \text{si } y_i = -1$$

És a dir, $C(x) = \text{sign}(f(x))$ ens classifica els punts per classes.

Vista la descripció geomètrica del cas linealment separable, comencem amb l'optimització del marge.

3.3.2 El problema d'optimització de la LSVM

Suposem que les nostres classes són linealment separables, i per això, podem trobar una funció $f(x) = \beta^t x + \beta_0$ tal que, $y_i f(x_i) \geq 0$, $\forall i$.

Si busquem el marge màxim que separi les classes, tenim el següent problema d'optimització (problema *primal*):

$$\max_{\beta, \beta_0, \|\beta\|=1} M, \tag{3.7}$$

$$\text{subjecte a les condicions: } y_i(\beta^t \cdot x_i + \beta_0) \geq M, \, \forall i \in [1, n].$$

El conjunt de condicions, ens confirmen que tots els punts es troben a una distància superior a M de la frontera de decisió definida per β i β_0.

Reparametritzem el problema, introduint $\|\beta\|$ dins de la condició.

$$y_i(\beta^t \cdot x_i + \beta_0) \geq M\|\beta\|$$

Assignem arbitràriament $\|\beta\| = 1/M$, ja que per a qualsevol múltiple de β i β_0 la desigualtat es compleix.

$$\min_{\beta, \beta_0} \frac{1}{2}\|\beta\|^2, \tag{3.8}$$

$$\text{subjecte a les condicions: } y_i(\beta^t \cdot x_i + \beta_0) \geq 1, \forall i \in [1, n].$$

Definició 7 (Multiplicador de Lagrange)**.** *Per a resoldre el problema de maximitzar una funció $f(x, y)$ subjecte a una condició $g(x, y) = c$, es considera la funció Lagrangiana:*

$$\mathcal{L}(x, y, \lambda) = f(x, y) - \lambda(g(x, y) - c),$$

on λ és un paràmetre, el multiplicador de Lagrange.

Problema *primal* d'optimització

Apliquem, la *funció Lagrangiana* per resoldre el problema *primal* d'optimització, trobar el mínim de la funció depenent dels paràmetres β i β_0 :

$$\mathcal{L}(\beta, \beta_0, \alpha) = \frac{1}{2}\|\beta\|^2 - \sum_{i=1}^{n} \alpha_i[y_i(\beta^t x_i + \beta_0) - 1], \tag{3.9}$$

on $\alpha = \{\alpha_i\}_{1 \leq i \leq n}$ són multiplicadors.

Fem les derivades parcials igualant a 0.

$$\frac{\partial L}{\partial \beta}(\beta, \beta_0, \alpha) = 0, \quad \frac{\partial L}{\partial \beta_0}(\beta, \beta_0, \alpha) = 0.$$

Obtenim,

$$\beta = \sum_{i=1}^{n} \alpha_i y_i x_i, \quad \sum_{i=1}^{n} \alpha_i y_i = 0. \tag{3.10}$$

Problema dual

Tornem a sustituir a (3.9) els resultats anteriors, trobant-nos amb el problema dual:

$$Q(\alpha) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{k=1}^{n} \alpha_i \alpha_k y_i y_k {}_i^t x_k, \tag{3.11}$$

$\alpha_i \geq 0$, $1 \leq i \leq n$

Aquest cop, volem maximitzar $Q(\alpha)$. La solució ha de complir les condicions de *Karush-Kuhn-Tucker*, incloent les (3.10), (3.11) i

$$\alpha_i[y_i(\beta^t x_i + \beta_0) - 1] = 0 \ \ \forall i \in [1, n]. \tag{3.12}$$

D'aquesta darrera fórmula, podem entendre que:

- si $\alpha_i > 0$, x_i es troben sobre el marge màxim.
- si $y_i(\beta^t x_i + \beta_0) > 1$, x_i no és al marge màxim i $\alpha_i = 0$.

Els punts que es troben en el marge màxim, s'anomenen *punts suport*.

Amb la resolució de (3.10) i (3.12), un cop trobat l'hiperplà amb marge màxim,

$$\hat{f}(x) = \hat{\beta}^t x + \hat{\beta}_0$$

es defineix $\hat{C}(x) = \text{sign}(\hat{f}(x))$ per a classificar les noves observacions.

3.3.3 Classificador LSVM

Un cop vist LSVM, farem algunes suposicions que afectaran a la separabilitat lineal, i així introduirem els SVC (*Support Vector Classifiers*).

Suposem ara, que el nostre conjunt de dades no és linealment separable, és a dir, existeixen alguns punts que es troben mal classificats, com *outliers*, en la banda incorrecte del marge.

Definim les variables de fluixedat marginal $\epsilon = (\epsilon_1, \epsilon_2, \ldots, \epsilon_n)$. Afegint aquesta nova variable al problema de separació, modifiquem la condició de (3.7) :

$$y_i(\beta^t x_i + \beta_0) \geq M(1 - \epsilon_i), \forall i, \tag{3.13}$$

$\epsilon_i \leq 0$, $\sum_{i=1}^n \epsilon_i \leq K$, constant.

Sabent que ϵ_i a (3.13) és proporcional a la mala classificació de l'observació. Acotant $\sum_{i=1}^n \epsilon_i$, aconseguim fitar la quantitat total de punts malaments classificats. La mala classificació passa quan $\epsilon_i > 1$, acotant $\sum_{i=1}^n \epsilon_i$ per un valor K, acotem el total de mala classificació amb K.

Com en el cas de separació lineal, per a trobar la solució hem de resoldre:

$$\min\|\beta\| \text{ subjecte a les condicions: } \begin{cases} y_i(\beta^t \cdot x_i + \beta_0) \geq 1 - \epsilon_i, \forall i, \\ \epsilon_i \geq 0, \sum \epsilon_i \leq \text{constant.} \end{cases} \tag{3.14}$$

Reajustem (3.14), per plantejar el problema d'optimització *primal*.

$$\min_{\beta,\beta_0} \|\beta\|^2 + \gamma \sum_{i=1}^n \epsilon_i$$

$$\text{subjecte a les condicions: } \epsilon_i \geq 0,\ y_i(\beta^t x_i + \beta_0) \geq 1 - \epsilon_i,\ \forall i$$

γ sustitueix la constant a (3.14).

Problema *primal* d'optimització

Resolem el problema *primal* d'optimització, minimitzant la següent fórmula respecte β, β_0 i ϵ_i :

$$\mathcal{L}(\beta, \beta_0, \alpha) = \frac{1}{2}\|\beta\|^2 + \gamma \sum_{i=1}^n \epsilon_i - \sum_{i=1}^n \alpha_i[y_i(\beta^t x_i + \beta_0) - (1 - \epsilon_i)] - \sum_{i=1}^n \mu_i \epsilon_i \tag{3.15}$$

Sustituint les derivades parcials,

$$\beta = \sum_{i=1}^{n} \alpha_i y_i x_i, \tag{3.16}$$

$$0 = \sum_{i=1}^{n} \alpha_i y_i, \tag{3.17}$$

$$\alpha_i = \gamma - \mu_i, \forall i, \tag{3.18}$$

en (3.15), obtenim el corresponent problema dual,

$$Q(\alpha) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{k=1}^{n} \alpha_i \alpha_k y_i y_k {}_i^t x_k,$$

Problema dual

La solució maximal de $Q(\alpha)$ restringida a $0 \leq \alpha_i \leq \gamma$ i $\sum_{i=1}^{n} \alpha_i y_i = 0$. Juntament restringida amb les condicions de *Karush-Kuhn-Tucker* que inclouen:

$$\alpha_i[y_i(\beta^t x_i + \beta_0) - 1 - \epsilon_i)] = 0, \tag{3.19}$$

$$\mu_i \epsilon_i = 0, \tag{3.20}$$

$$y_i(\beta^t \beta + \beta_0) - (1 - \epsilon_i) \geq 0, \tag{3.21}$$

per a $i \in [1, n]$. El conjunt d'equacions que trobem, des del càlcul de les derivades parcials fins ara, caracteritzen de manera única el problema *primal* i el *dual*.

Veiem, per (3.16) que la solució de β és de la forma:

$$\hat{\beta} = \sum_{i=1}^{n} \hat{\alpha}_i y_i x_i,$$

amb $\hat{\alpha}_i \neq 0$ per a les observacions i que compleixen (3.17) i (3.18). Aquestes observacions s'anomenen *vectors suport*. Per a més informació, consultar [1].

Donades les solucions $\hat{\beta}_0$ i $\hat{\beta}$, la funció de classificació la escrivim així:

$$\hat{C}(x) = \text{sign}(\hat{f}(x)) = \text{sign}(\hat{\beta}^t x + \hat{\beta}_0)$$

El paràmetre que diferencia aquest procediment, del que no té ϵ és γ.

3.3.4 SVM no lineal

En aquest apartat, tractarem amb dades no separables i veurem com es soluciona aquest problema sense caure en el sobreajustament.

Com hem introduït abans, de vegades ens trobem amb problemes no lineals a $\mathbb{T}$, però transformant-los en un espai adequat, $\mathbb{X}$, els sabem resoldre. Estem parlant de l'ús de *kernels*, nuclis.

Definició 8. *La transformació que linealitzarà el problema*

$$\Phi : \mathbb{T} \to \mathbb{X}$$

$$t \mapsto x,$$

queda determinada a partir d'una funció de dues variables, $K(t_1, t_2)$ *que s'anomena nucli.*

Sigui $(x_i, y_i), \ldots, (x_n, y_n)$ un conjunt d'aprenentatge, amb $\{x_i\} \in \mathbb{R}^p$ i p variables predictores. S'aconsegueix linealitzar els problemes que no ho són, obligant que els càlculs només depenguin dels n^2 productes escalars $\langle x_i, x_j \rangle$.

Definim, la funció nucli $K(\cdot, \cdot)$ de manera que:

$$\langle x_i, x_j \rangle := K(t_i, t_j),\ 1 \leq i,\ j \leq n,$$

mantenint les propietats dels n^2 productes de x_i, amb $K(\cdot, \cdot)$ una funció definida positiva, és a dir, $\forall t_1, \ldots, t_n \in \mathbb{T}$, $(K(t_i, t_j)) \geq 0$.

Sustituint aquest resultat a la funció de Lagrange Dual (3.11), obtenim:

$$Q(\alpha) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{k=1}^{n} \alpha_i \alpha_k y_i y_k \langle \Phi(t_i), \Phi(t_k) \rangle ,$$

$\alpha_i \geq 0, 1 \leq i \leq n$.

Anàlogament, fent ús de (3.10) i aplicant els kernels a la funció solució obtenim:

$$f(x) = \Phi(t)^t \beta + \beta_0 = \sum_{i=1}^{n} \alpha_i y_i \langle \Phi(t), \Phi(t_i) \rangle + \beta_0$$

Els nuclis més populars són:

- Nucli polinomial de grau d: $K(x_i, x_j) = (\langle x_i, x_j \rangle + 1)^d$
- Nucli amb base radial: $K(x_i, x_j) = exp(-\|x_i - x_j\|^2/c)$
- Nucli sigmoidal: $K(x_i, x_j) = tanh(a \langle x_i, x_j \rangle + b)$

Per acabar la solució seria de la forma:

$$\hat{f}(x) = \sum_{i=1}^{n} \hat{\alpha}_i y_i K(x, x_i) + \hat{\beta}_0$$

Consultar [1] per a més informació.

4 Descripció de les dades

En aquest apartat, es pretén formar una idea de la distribució de la mostra que tractem. Inicialment, el fitxer contenia un conjunt de 117205 observacions amb 125 variables predictores, que englobaven tot tipus de dades. Un cop efectuats els primers passos del preprocessament trivials com l'eliminació de columnes buides, duplicats, entre d'altres, el conjunt resultant es mostra a continuació:

Nom de la variable	**Nivells**	**Abreviació**
CATEGORIA_TRANSMITENT	**1** Administració	**1** AD
	2 Altres	**2** AL
	3 Financera	**3** FI
	4 Immobiliària	**4** IM
COD_MODEL	**1** 600	
	2 651	
COD_TARIFA	**1** Model 651	**1** M651
	2 + Actes jurídiques	**2** RAJ
	3 Acta jurídica 1	**3** AJ1
	4 Acta jurídica 2	**4** AJ2
	5 Acta jurídica 5	**5** AJ5
	6 Transmissions patrimonials	**6** TRT
	7 + Transmissions patrimonials	**7** RTP
EXERCICI	{1978,1999-2014}	
ESTAT_EXPEDIENT	**1** FI	
	2 PC	
	3 TE	
TIPO_SUBTIPO_EXPE	**1** Autoliquidació de Donacions	**1** SU.A1
	2 Expedient d' ofici de Donacions	**2** SU.E1
	3 Model 600	**3** TR.00
FASE_1	**1** DEPURAT	
	2 SUPERFICIE FORA MARGE	
	3 VALOR TOTAL FORA MARGE	
FASE_2	**1** DEPURAT	
	2 MANCA INFORMACIÓ SUPERFICIE	
FASE_3	**1** DEPURAT	
	2 VALOR CNP FORA MARGE	
TIPUS_SUBTIPUS_BIEN	**1** 03.IU.HP	
	2 21.IM.HP	

TIPUS_SITUACIO	**1** A **2** B **3** C	
TIPOLOGIA_OMIC	**1** Altres **2** Valors de repercussió ús habitacle	**1** AL **2** VH
CLAU_US_CADASTRE	12 Nivells	
CATEGORIA	**1** Grup inexistent **2** Bons **3** Mitjans **4** Dolents **5** Grup ABC	**1** GI **2** B **3** M **4** D **5** GABC
TRAMS_ANTIGUITAT	**1** 0-5 **2** 6-10 **3** 11-14 **4** 15-29 **5** 30-49 **6** 50-69 **7** 70-99 **8** +100	
COEF_INSTRUCCIO		
COEFICIENT_PROPIETAT		
POBLACIO		
PERIODE		
COEF_CADASTRAL_DECLARAT		
VALOR_CNP_GAUDI		
VALOR_CNP_CALCULAT		
VALOR_MOSTRA		
VALOR_MOSTRA_DECLARAT		
VALOR_PARTICIPACIO_DECLARADA		
VALOR_MOSTRA_PROJECTAT_ANY0		
VALOR_TOTAL_MAXIM		
ANTIGUITAT		
SUPERFICIE_CONSTRUIDA		
SUPERFICIE_SOL		
VALOR_CADASTRAL		

Taula 1: Variables de la base de dades

Vista la taula anterior, s'afirma que les variables es reparteixen en 15 categòriques i 22 contínues. Seguidament, per a tenir una idea més exhaustiva dels valors que prenen cadascuna d'aquestes variables es presenta un resum estadístic de les categòriques que disposem:

CATEGORIA_TR	COD_MODEL	COD_TARIFA	EXERCICI	ESTAT_EXP	TIPO_SUBTIPO_EXPE
1: 1793	**1**: 34496	**1**: 1368	**2014**: 83	**1**: 34539	**1**: 1357
2: 31078	**2**: 1368	**2**: 5769	**2013**: 1765	**2**: 1321	**2**: 11
3: 1930		**3**: 203	**2012**: 4375	**3**: 4	**3**: 34496
4: 1063		**4**: 209	**2011**: 4286		
		5: 179	**2010**: 4359		
		6: 2	**2009**: 4051		
		7: 28134	**2008**: 3955		
			2007: 4655		
			2006: 5323		
			2005: 2966		
			2004: 46		

FASE_1	FASE_2	FASE_3	TIPUS_SUBTIPUS_BIEN	TIPUS_SITUACIO	CATEGORIA
1: 30540	**1**: 35834	**1**: 14837	**1**: 1368	**1**: 21027	**1**: 6
2: 5233	**2**: 30	**2**: 21027	**2**: 34496	**2**: 65	**2**: 824
3: 91				**3**: 14772	**3**: 33842
					4: 1191
					5: 1

CLAU_US_CADASTRE	TIPOLOGIA_OMIC	TRAMS_ANTIGUITAT
1: 34176	**1**: 4215	**1**: 14837
2: 535	**2**: 31649	**2**: 21027
3: 504		
4: 228		
5: 218		
6: 147		
7: 56		

Taula 2: Taula de freqüències per a les variables categòriques

De les variables categòriques contingudes a les dades, la majoria d'elles no superen els 4 nivells, cal destacar que algunes de les variables: CATEGORIA, ESTAT_EXP, TIPO_SUBTIPO_EXPE tenen poques observacions en alguns dels nivells i això s'haurà de tenir en compte posteriorment.

A continuació, es presenta un resum estadístic per a les variables contínues.

Estadídistic	Mitjana	St. Dev.	Min	Max
VARIABLE_RESPOSTA	0.024	0.181	−2.039	0.998
COEF_INSTRUCCIO	2.687	0.917	1.000	6.600
COEFICIENT_PROPIETAT	12.758	23.915	0.000	100.000
POBLACIO	266,880.900	502,981.600	105	1,611,822
PERIODE	0.482	0.290	0.000	1.000
COEF_CADASTRAL_DECLARAT	3.573	6.531	0.000	678.860
RM_DECLARAT	0.628	1.747	0.000	100.000
SUPERFICIE_DECLARADA	101.179	331.126	0.000	51,380.000
COEF_DISPERSIO	0.286	0.359	0.000	20.670
COEF_DISPERSIO_DECLARAT	0.335	0.356	0.000	21.890
VALOR_CNP_GAUDI	114,656.400	181,483.200	0.000	11,071,075.000
VALOR_CNP_CALCULAT	171,048.400	371,121.000	2,581.710	17,915,587.000
VALOR_MOSTRA	1,767.626	1,066.955	0.000	40,246.360
VALOR_MOSTRA_DECLARAT	1,609.860	1,090.105	0.000	40,246.360
VALOR_PARTICIPACIO_DECLARADA	134,477.800	137,526.900	450.240	7,000,000.000
VALOR_MOSTRA_PROJECTAT_ANY0	1,646.511	1,427.624	0.000	66,878.310
VALOR_MOSTRA_PROJEC_DECLA_ANY0	1,508.342	1,403.737	0.000	65,038.120
VALOR_TOTAL_MAXIM	171,257.100	170,771.800	2,204.280	9,515,977.000
ANTIGUITAT	32.912	29.846	0	1,802
SUPERFICIE_CONSTRUIDA	126.972	244.080	0	11,784
SUPERFICIE_SOL	1,206.756	2,211.367	0	80,602
VALOR_CADASTRAL	71,826.710	148,286.800	956.190	5,598,621.000

Taula 3: Resum estadístic per a les variables numèriques

De la taula anterior, el primer punt a destacar és la dispersió que existeix entre algunes variables, com POBLACIO, així com la diferència de rangs entre les variables, p.e COEF_DISPERSIO i VARLOR_TOTAL_MAXIM. Amb això es pot afirmar que les distribucions de les dades seran variants on algunes tindran els registres molt concentrats en una franja i en d'altres més dispersos.

5 Metodologia

En aquest apartat s'exposen en ordre d'execució els diferents processos que han sofert les dades, des de l'inici, de l'extracció de les dades del fitxer fins a les taules de resultats.

5.1 Preprocessat

La taula 4 dóna una idea per saber com era el conjunt de dades inicial abans de ser tractat i resultar la taula 1 on s'han implementat els diferents mètodes. L'objectiu d'aquesta etapa és arribar a obtenir el número de variables i d'observacions mínim que faci òptima la predicció sense caure en el sobre ajustament.

L'elecció de les diferents variables, així com també l'eliminació d'algunes observacions, i el seu tractament previ s'ha realitzat en una sèrie de passos:

Fases aplicades a la taula original, Taula 4

Num.Observacions	Num.Variables Categòriques	Num.Variables Numèriques
117205	41	84

Taula 4: Estructura de les dades inicials

- Eliminació de variables buides, quasi buides, d'un nivell i d'observacions duplicades.
- Tractament d'*outliers*: aplicació de la distància de Cook, per detectar les observacions amb residus tan grans que distorsionarien la predicció.
- Filtres funcionals: s'apliquen diferents restriccions sobre les dades segons la demanda de l'expert funcional, concretament, el cap del projecte que sàpiga els objectius d'aquest, eliminació de variables amb informació irrellevant així com també creació de noves variables a partir d'altres ja existents per optimitzar el propòsit de la predicció. *Ex: imposar dominis a les variables, selecció de nivells específics, creació de la variable resposta, etc.* En aquest cas, s'han reduït les observacions a les que corresponen a immobles personals urbans.
- Anàlisi estadística: estudi de les correlacions entre les variables numèriques, així com també entre la variable resposta i les variables categòriques per saber el seu nivell de significança. *Ex: eliminació de variables generades per altres ja existents.*

  Comparant la gràfica (2), als Annexos, de les dades actuals amb la gràfica (9) de correlacions (9) de la taula inicial, es pot veure la diferència del número de variables correlacionades.

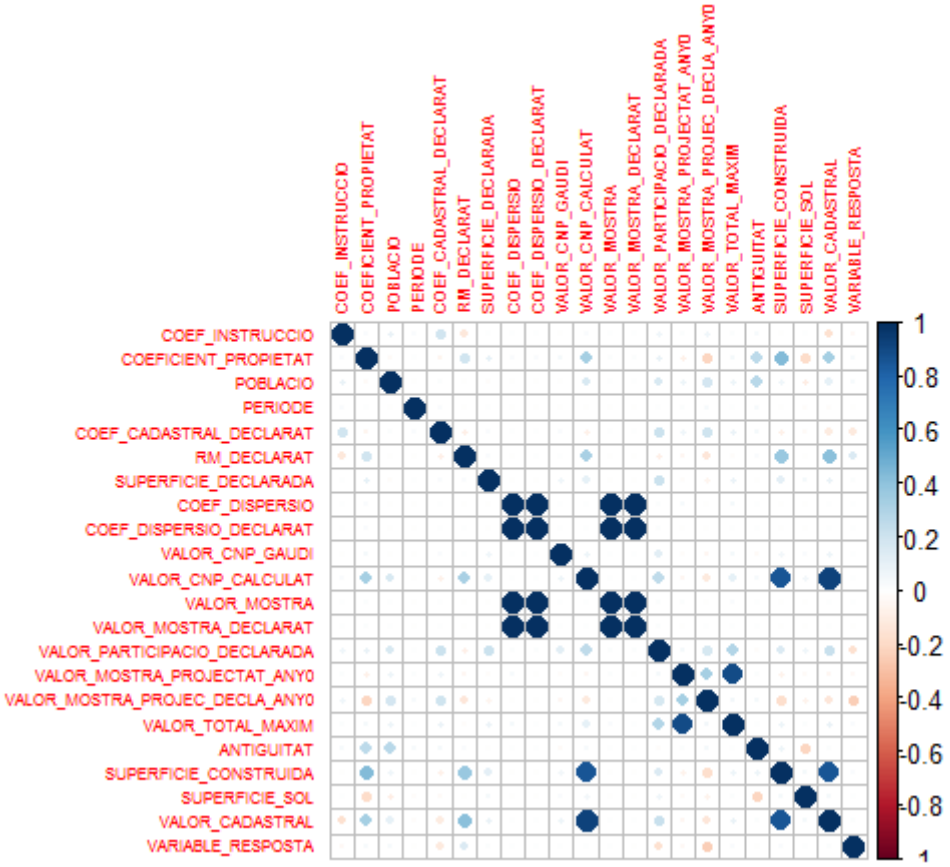

Figura 2: Correlació entre les variables numèriques

Fases aplicades a la taula ajustada, Taula 1

- Anàlisi descriptiva: mesures de dispersió, de posició i de centralització. *Ex: Re configuració dels nivells amb freqüències baixes.*.

  Com s'ha vist en la secció anterior, hi havien nivells d'algunes variables amb molt poques observacions; per a evitar problemes d'aparició de nivells nous a l'hora de testejar, s'han reagrupat nivells.

- Estudi de la normalitat, fent ús de proves gràfiques.

- Centralització i normalització de les dades, per a que cada variable tingui mitjana 0 i desviació estàndard 1.

  Abans, a la descripció de les dades, s'ha esmentat el fet que hi havia variables amb rangs molt variables. Les estandaritzem, per aconseguir millors resultats.

Un error que es pot prevenir amb l'escalabilitat de les dades és el *rank-deficient*. Aquest es pot generar per diferents motius, tots relacionats amb el rang de la matriu de dades, com ara multicolinealitat, freqüències baixes en nivells de variables categòriques o l'existència de variables constants.

5.2 Anàlisi dels resultats

Posteriorment al preprocessat, es comença amb la implementació de cadascun dels mètodes anteriorment citats.

Mesura de qualitat

En les subseccions següents es presentaran l'anàlisi dels models i la seva validació. Per poder fer els resultats i la discussió d'aquests, es necessita tenir un conjunt comú d'entrenament i de test, per aquesta raó s'han dividit les dades en dues parts:

- 2/3 parts de les dades totals s'assignen a l'entrenament dels models seleccionats, reajustant els paràmetres necessaris.
- 1/3, part restant de les dades totals, correspondrà a l'utilitzat per al testeig i validació de la capacitat de predicció d'aquests.

Hem d'enfatitzar que es vol optimitzar, reduint el número de *Falsos Negatius*, FN, sense oblidar el percentatge d'error.

Notació: F = Frau , NF = No Frau

Alguns dels indicadors que utilitzarem per analitzar els diferents resultats de les classificacions seran les matrius de confusió i les taules dels *ratios* de la forma següent:

Ratio TP (sensibilitat o *recall*)	$\frac{TP}{TP+FN}$	Ratio FN	$\frac{FN}{TP+FN}$
Ratio TN (especificació)	$\frac{TN}{TP+FN}$	Ratio FP	$\frac{FP}{TN+FP}$
	accuracy[4]	$\frac{TP+TN}{TP+TN+FP+FN}$	

Taula 5: Taula dels *ratios*

[4] Proporció (tant per u) de classificacions correctes al conjunt de prova.
TP = *True Positive*, FP = *False Positive*,
FN = *False Negative* i TN = *True Negative*.

Definició 9. *Anomenem precisió* (accuracy) *una estimació de la probabilitat (global) de no cometre error de classificació.*

5.2.1 Naive Bayes

En aquest apartat, s'analitza el primer model dels tres que s'han seleccionat, el classificador Naive Bayes. L'objectiu d'aquest apartat és veure quines estimacions ens proporciona Naive Bayes.

Creem un model Naive Bayes amb les dades del conjunt d'entrenament que tenim:

		Predicció		
		F	NF	**Totals**
Reals	F	1437	293	1730
	NF	1063	1989	3052
	Totals	2500	2282	4782

Taula 6: Matriu de confusió

A partir de la matriu de confusió obtenim la següent taula dels *ratios*:

Ratio TP (sensibilitat o *recall*)	0.8306	Ratio FN	**0.1693**
Ratio TN (especificació)	0.6517	Ratio FP	0.3482
	accuracy	0.7164	

Taula 7: Taula dels *ratios*

Com podem veure en aquesta darrera taula, el percentatge d'observacions ben classificades ha sigut d'un 71% i apart, el *ratio* dels *Fals Negatius* ha sigut bastant baix.

Per assegurar-nos que aquest comportament no és excepcional, sinò que es troba dins del rang d'estimació normal d'aquest model, hem implementat un mètode k-folds. Aquest mètode consisteix en dividir el conjunt de dades en k mostres diferents amb la mateixa dimensió i de manera recursiva, anar entrenant el model amb k-1 de les k mostres i la restant, utilitzar-la com test per poder estimar l'eficàcia del mètode. D'aquesta forma, obtenim k estimacions que al fer el promig, ens diu si la predicció de la nostra mostra puntual és anormal o s'ajusta a les prediccions esperades.

Hem comprovat com es comporta Naive Bayes, aplicant k-folds amb $k = 10$ i efectivament, ha retornat una precisió del 71%, per tant no tenim un model inestable.

5.2.2 Arbres CART i *Bagging*

Com hem vist a l'apartat d'aproximació teòrica, el mètode d'ensemble *bagging* normalment utilitza mètodes inestables, ja que una de les característiques d'aquest és la reducció de la variància.

Anem a comprovar la diferència en l'estimació de la variable resposta, primer amb arbres CART de classificació sense *bagging* i *a posteriori* amb *bagging*.

A continuació, es mostra l'arbre generat a partir de les dades d'entrenament (Fig.3).

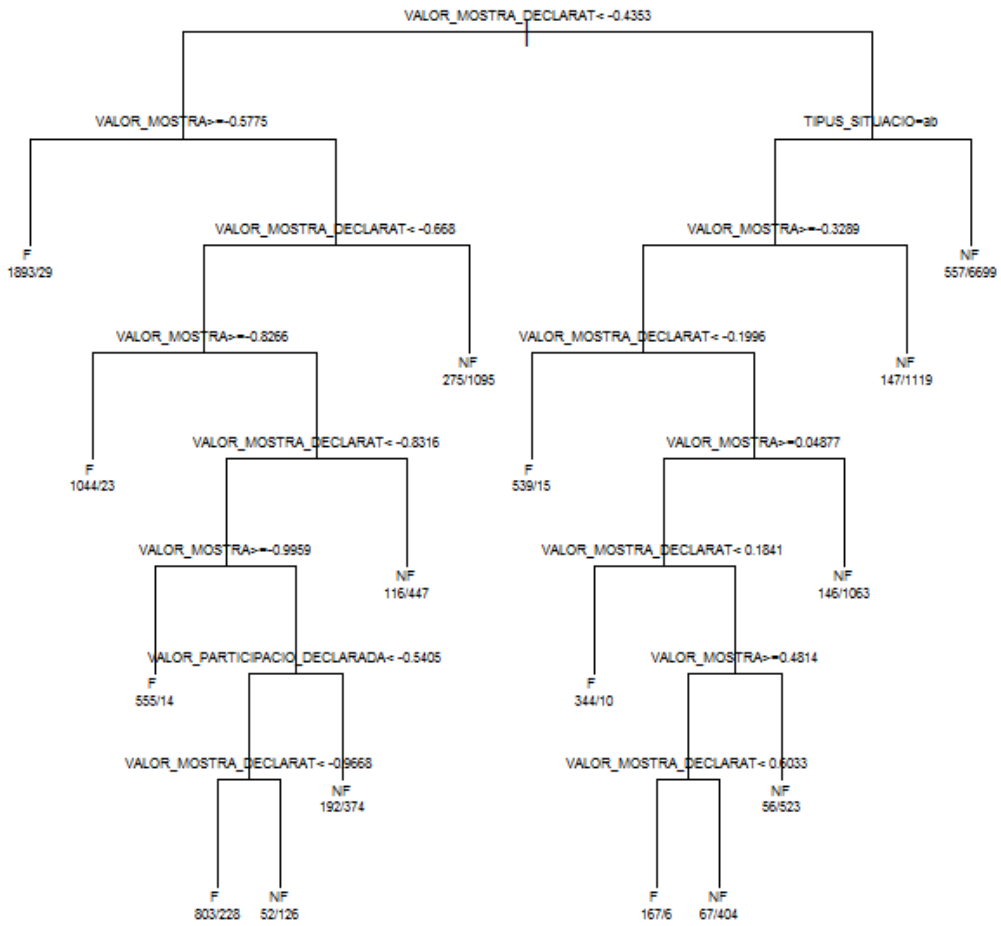

Figura 3: Arbre de classificació

Un cop vist el càlcul, es vol aconseguir l'arbre òptim, podant l'arbre amb el CP òptim. Aquest valor CP, correspòn al *complexity parameter* esmentat en la secció d'aproximació teòrica.

El criteri que normalment es segueix és escollir l'arbre que minimitza l'error de la validació creuada. Aquests valors els podem veure a la columna *xerror* de la taula 8 i en el gràfic següents:

	CP	nsplit	rel error	xerror	xstd
1	0.3730764	0	1.0000000	1.0000000	0.0095678
2	0.0589674	1	0.6269236	0.6298001	0.0083572
3	0.0251211	3	0.5089889	0.5144542	0.0077559
4	0.0240184	6	0.4336258	0.4304617	0.0072265
5	0.0238027	8	0.3855890	0.4179491	0.0071399
6	0.0130879	10	0.3379836	0.3540918	0.0066612
7	0.0115777	12	0.3118079	0.3303610	0.0064659
8	0.0106429	14	0.2886524	0.3172731	0.0063536
9	0.0100000	15	0.2780095	0.3132461	0.0063184

Taula 8: CP en funció del número de branques

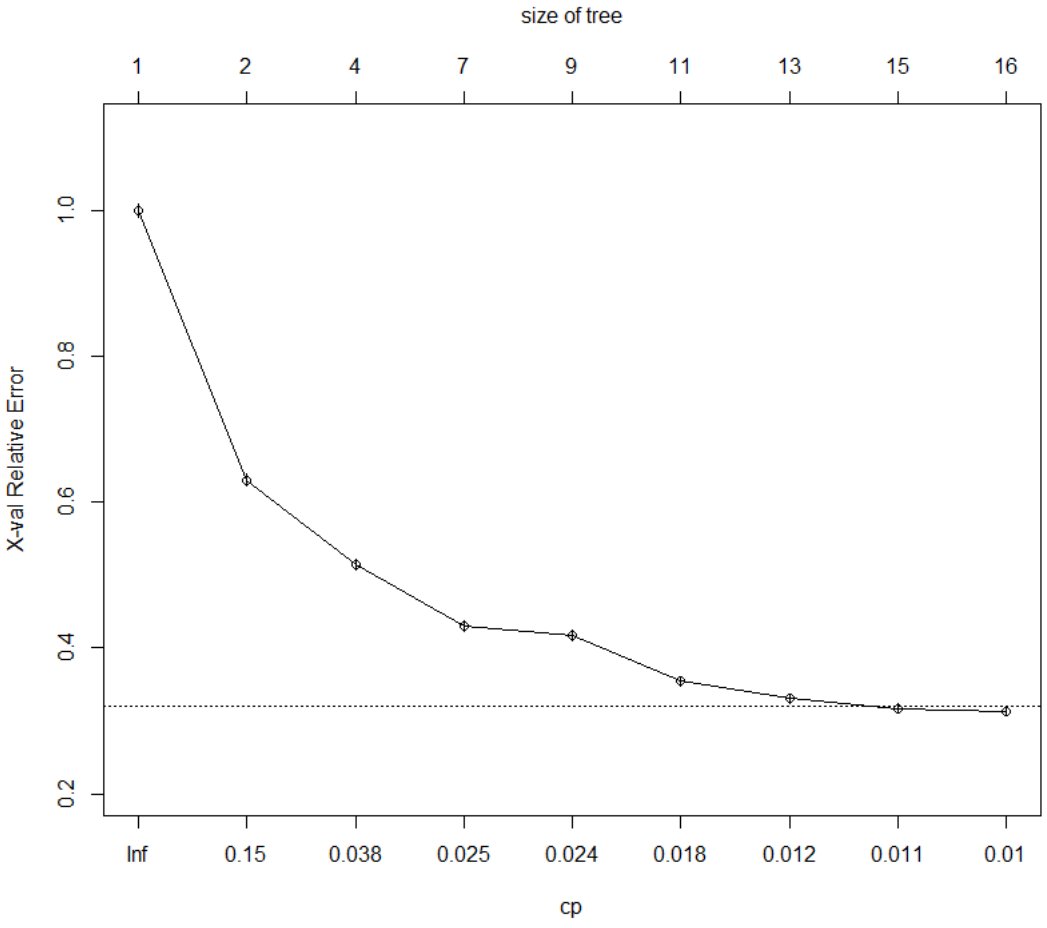

Figura 4: Errors de la validació creuada segons el tamany de l'arbre

A partir de la taula i del gràfic podem veure que els millors resultats s'obtenen per un CP equivalent a 0,01. Així, escollim aquest valor com a òptim i l'utilitzem per a recalcular l'arbre anterior. L'arbre òptim coincideix amb l'arbre màxim, per aquesta raó si apliquem la poda a l'arbre inicial obtindrem el mateix.

L'arbre definitiu és:

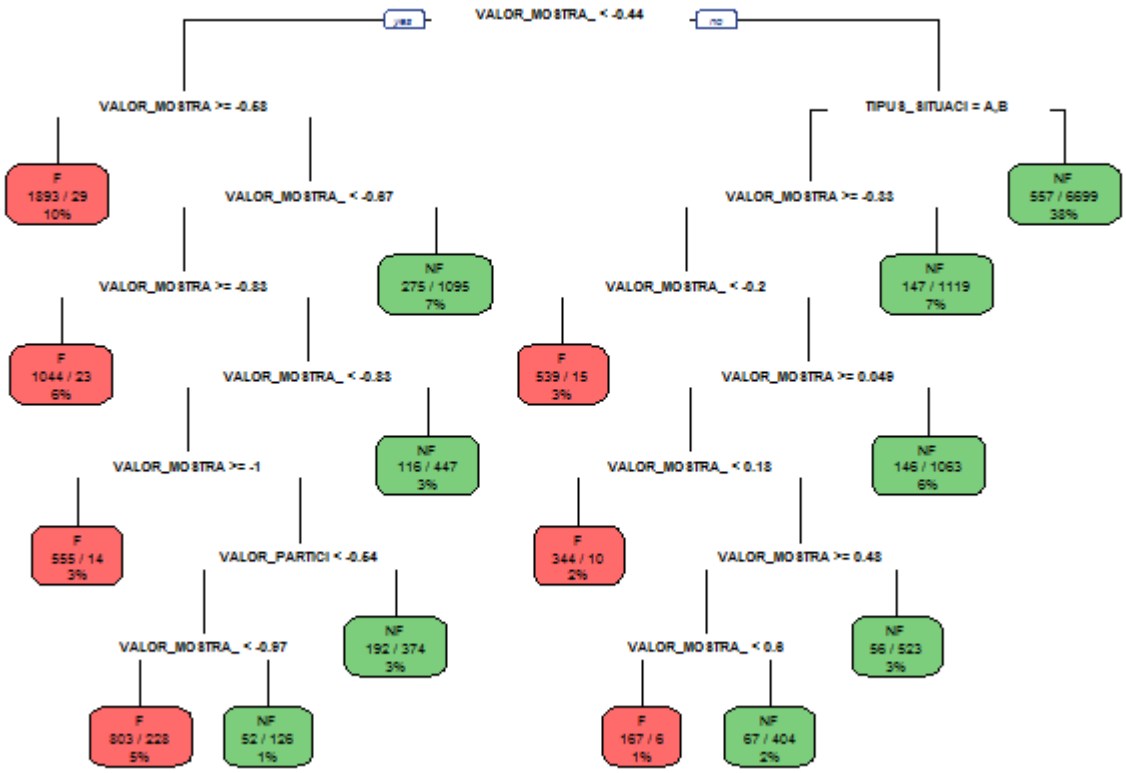

Figura 5: Arbre de classificació òptim

Com podem veure, les variables que s'han utilitzat com a nodes en aquest arbre són: TIPUS_SITUACIO, VALOR_MOSTRA, VALOR_MOSTRA_DECLARAT i VALOR_PARTICIPACIO_DECLARADA. A més d'això, podem observar quants elements malament classificats es troben a les diferents fulles. El total d'observacions mal classificades és de 2000, que correspon a un 9,56% del conjunt d'entrenament. Apart d'això, a cada node també podem veure el percentatge de dades que conté.

Ara anem a analitzar com són les seves estimacions. Comencem amb la matriu de confusió que generem a partir de les prediccions fetes amb el conjunt *test*:

		Predicció		
		F	NF	**Totals**
Reals	F	1331	399	1730
	NF	99	2953	3052
	Totals	1430	3352	4782

Taula 9: Matriu de confusió

Partint de la matriu de confusió obtenim la següent taula dels *ratios*:

Ratio TP (sensibilitat o *recall*)	0.7693	Ratio FN	**0.2306**
Ratio TN (especificació)	0.9675	Ratio FP	0.0324
	accuracy	0.8958	

Taula 10: Taula dels *ratios*

Com observem a les taules 10 i 9, els arbres de classificació respecte el classificador Naive Bayes, tot i mostrar una millor precisió, el percentatge de *Falsos Negatius* és més elevat als arbres de classificació.

A continuació, hem volgut veure com d'inestable era l'arbre CART, fent un k-folds amb $k = 10$.

iteració	*accuracy*
1	0.8834
2	0.9007
3	0.8887
4	0.8973
5	0.8954
6	0.8818
7	0.9072
8	0.8959
9	0.8960
10	0.8873
Promig	0.8933

Taula 11: Arbres CART de classificació amb k-folds, $k = 10$

Mirant el resultat de la taula on hem aplicat el k-folds, estem comprovant que realment aquest arbre no és tan inestable com esperàvem ja que té una diferència molt petita respecte al resultat de la taula 9 i també, si mirem les diferents estimacions pels diferents k, no hi ha molta variació.

Apliquem el *bagging* als arbres i mirem quina és la seva acuracitat segons el número de mostres que genera:

	#bosses	*accuracy*	Ratio FN
1	20	0.9548	0.0849
2	21	0.9601	0.0774
3	22	0.9590	0.0791
4	23	0.9588	0.0797
5	24	0.9607	0.0803
6	25	0.9592	0.0815

Taula 12: Relació d'encertats respecte el nombre de mostres *bootstrap*

Si mirem la taula anterior, primerament podem observar que la precisió ha augmentat respecte a la que ens retornàven els arbres CART, és a dir, *bagging* ens ha millorat l'estimació de la variable resposta. En segon lloc, podem afirmar que dins del conjunt de bosses escollit, el resultat més òptim amb major nombre d'encerts i menys *Falsos Negatius* el trobem agafant 21 mostres *bootstrap*.

5.2.3 SVM

En aquest apartat es crea un tercer model per al càlcul de prediccions, fent ús de SVM (*Support Vector Machines*).

Per calcular quins són els paràmetres que s'ajusten millor a les dades amb les que treballem, provem amb diferents γ i diferents costs, fent ús del *grid-search*.

Comencem amb un domini menys acurat, amb combinacions d'intervals de γ de 2^{-15} fins 2^5 (amb increments en l'exponent de 2.5) i de costs de 2^{-5} fins 2^{10} (amb increments en l'exponent de 2.5)

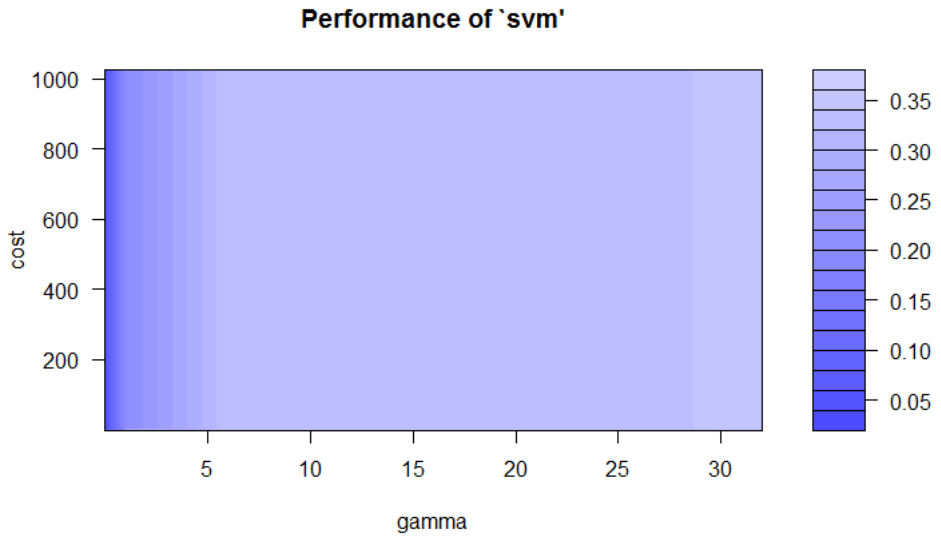

Figura 6: Valors de gamma de 2^{-15} fins 2^5 i costs de 2^{-5} fins 2^{10}

En la figura (6), podem observar que els millors resultats es troben quan la γ pren valors més petits. Per a valors superiors a 2, veiem que els resultats obtinguts no són bones prediccions. Per altra banda, efectivament, la búsqueda de la parella de (γ,C) òptima és $(2^{-7.5}, 2^{10})$ amb un percentatge de precisió del 97.47%.

Ens aproximem als valors òptims de γ. Acurem l'interval de γ de 2^{-10} fins 2^{-5} (amb increments en l'exponent de 1) i l'interval de costs de 2^5 fins 2^{15} (amb increments en l'exponent de 2.5)

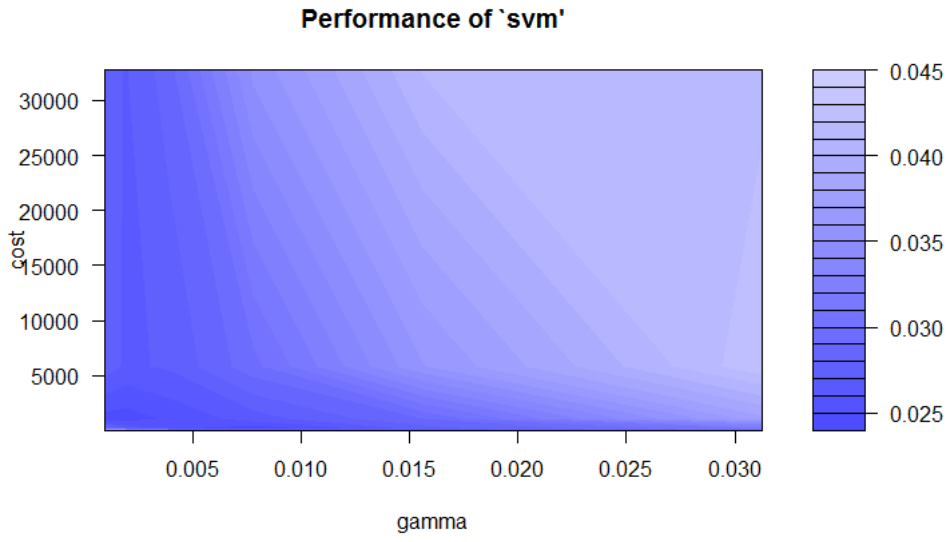

Figura 7: Valors de gamma de 2^{-10} fins 2^{-5} i costs de 2^{-5} fins 2^{15}

En la figura (7), es pot veure que els millors resultats es troben per a γ inferior a 2^{-6}, és a dir, inferior a 0.01. De fet, com hem escollit intervals més petits de γ i de cost, hem ajustat una mica el resultat a $(2^{-7}, 2^{7.5})$, com a valor òptims, amb un percentatge de precisió del 97.57%.

Un cop definit uns paràmetres de γ i cost adecuats, anem a modelar l'SVM amb aquests i veure quina estimació de *Falsos Negatius* ens retorna.

		Predicció		
		F	NF	**Totals**
Reals	F	1668	62	1730
	NF	57	2995	3052
	Totals	1725	3057	4782

Taula 13: Matriu de confusió

A partir de la matriu de confusió obtenim la següent taula dels *ratios*:

Ratio TP (sensibilitat o *recall*)	0.9641	Ratio FN	**0.03583**
Ratio TN (especificació)	0.9813	Ratio FP	0.01867
	accuracy	0.9751	

Taula 14: Taula dels *ratios*

Com a observació final, sense dubte l'SVM és el mètode que fa unes prediccions més acurades, amb un percentatge de *Falsos Negatius* més baix.

6 Tests

Optimitzats els paràmetres que proporcionen un dels millors resultats per cadascun dels tres mètodes escollits, estudiem quin d'ells ens retorna un nombre d'errors més petit i d'aquesta forma, seleccionar-lo com a model definitiu. Per realitzar aquesta decisió farem un k-folds amb $k = 10$ sobre els 2/3 de les dades assignades a l'aprenentatge dels models.

Per a cada iteració, entrenarem el model amb el 90% de les dades i el testejarem amb el 10% restant, calculant la precisió en la predicció.

Al final, compararem els diferents percentatges d'estimació correcta retornats per cadascun dels mètodes, per cadascun dels 10 grups de parelles test-entrenament.

En la taula 15 i en la gràfica 8 podem veure les estimacions correctes fetes pels diferents models:

iteració	Naive Bayes	*Bagging*	SVM
1	0.7163	0.9610	0.9764
2	0.7046	0.9549	0.9785
3	0.7110	0.9600	0.9798
4	0.6940	0.9666	0.9761
5	0.7022	0.9564	0.9729
6	0.7266	0.9574	0.9746
7	0.7040	0.9625	0.9716
8	0.7124	0.9547	0.9716
9	0.7038	0.9638	0.9815
10	0.7303	0.9539	0.9728
Promig	0.7105	0.9591	0.9755

Taula 15: Estimació de l'*accuracy* de cada model

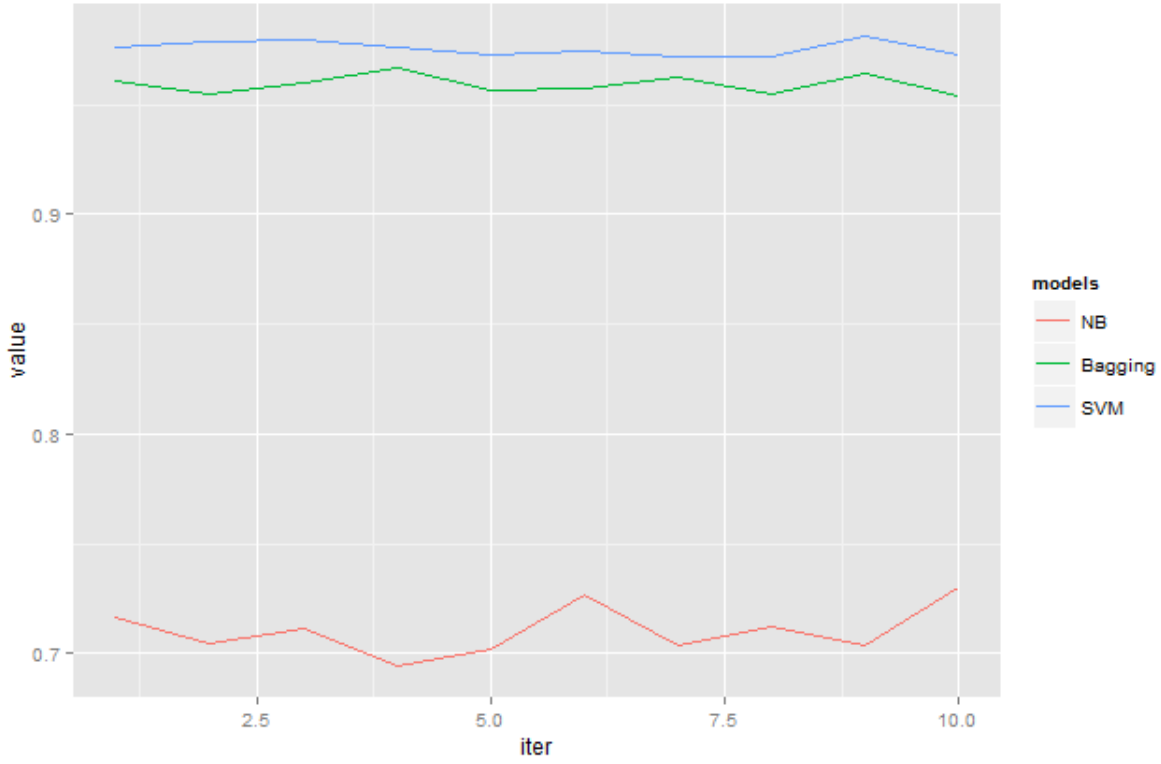

Figura 8: Precisió en cada iteració per als tres models

A partir del gràfic queda clarament reflexat que SVM és el model que proporciona uns resultats millors a l'hora d'estimar les classes. Tot i que, el mètode *bagging* també té unes prediccions bastants properes al SVM, o sigui bones.

Si ara mirem, la relació de *Falsos Negatius* de cadascun d'aquests mètodes, apreciarem una lleugera diferenciació entre *bagging* i SVM:

iteració	NB	Bagging	SVM
1	0.1681	0.0852	0.0384
2	0.1621	0.0907	0.0351
3	0.1970	0.0809	0.0352
4	0.1757	0.0674	0.0337
5	0.1633	0.0839	0.0447
6	0.1384	0.0860	0.0419
7	0.1371	0.0834	0.0480
8	0.1613	0.0870	0.0464
9	0.1936	0.0722	0.0355
10	0.1489	0.0902	0.0565
Promig	0.1645	0.0826	0.0415

Taula 16: FN generats a cada iteració pels models

El nombre de FN produïts per l'SVM és aproximadament la meitat dels ocasio-

nats pel *bagging*.

Així doncs, el model SVM serà l'escollit per fer les prediccions amb 1/3 de les dades que encara no han sigut utilitzades. A continuació, es mostra el resultat de la predicció:

		Predicció		
		F	NF	**Totals**
Reals	F	875	41	916
	NF	33	1441	1474
	Totals	908	1482	2390

Taula 17: Matriu de confusió

A partir de la matriu de confusió obtenim la següent taula dels *ratios*:

Ratio TP (sensibilitat o *recall*)	0.9552	Ratio FN	**0.0447**
Ratio TN (especificació)	0.97761	Ratio FP	0.0223
	accuracy	0.9690	

Taula 18: Taula dels *ratios*

Aquesta predicció, ens fa descartar la possibilitat de que el nostre model SVM estigués sobre ajustat, ja que si així fos, al fer prediccions sobre un conjunt de dades nou no hagués fet tan bona estimació.

7 Conclusions

Des de l'inici d'aquest treball, fins a arribar al final de la seva elaboració hem passat per vàries etapes complementàries entre elles, que han fet possible que al final tinguem una idea tant teòrica com pràctica, bastant completa, d'alguns dels mètodes de classificació estadística més utilitzats en l'actualitat.

Per començar, com es pot veure a la secció teòrica, Aproximació teòrica, ens hem introduït al món del *machine learning*, en mètodes d'aprenentatge automàtic supervisat, és a dir, aquells que necessiten d'un conjunt de dades d'entrenament per aprendre. Per fer la primera immersió en el tema, amb aquest projecte, ens vàrem centrar en mètodes de classificació, aquells llur predicció és de tipus qualitativa. Cadascun dels models, prové d'una branca de les matemàtiques diferent, tot i que compten amb una base comú estadística, per a tenir una mica de diversitat en l'aprenentatge. Els mètodes que hem estudiat i, posteriorment, aplicat són el classificador Naive Bayes, el mètode d'ensemble *bagging* aplicat a arbres de classificació CART i, finalment, les Màquines de Vectors Suport. De tots tres, hem seguit la seva fonamentació teòrica i, paral·lelament, hem dut a terme la seva implementació pràctica en conjunts de dades reals, comprovant la qualitat de les seves prediccions mitjançant l'interpretació de l'estimació de les probabilitats de cometre diversos tipus d'errors rellevants.

A l'hora d'aplicar aquestes tècniques de predicció, els coneixements teòrics adquirits amb anterioritat ens han ajudat a detectar alguns errors amb més facilitat. D'aquesta part més pràctica, hem de subratllar que el fet de treballar amb grans quantitats de dades en representació d'un problema real, fa que puguin sorgir incidències que normalment no es generen quan treballem amb un conjunt de dades tractat per ús acadèmic, per aquesta raó hem afegit una part de pre processat en el conjunt inicial de dades. Les dades emprades en aquest projecte, han sigut les proporcionades per l'empresa on estic cursant l'estada de pràctiques, per tant formen part d'un problema real que està en estudi: la detecció de transaccions de compra - venda fraudulentes.

Les estimacions obtingudes de l'aplicació pràctica dels models inclosos en el nostre pilot de mètodes, han concordat en general amb la idea que teníem sobre la qualitat que ens retornarien. El Naive Bayes ens ha retornat una precisió baixa respecte els altres dos (70% respecte 95-97%), ja que la presumpció d'independència entre les variables no sol ser correcta en dades reals, en el nostre cas particular, la rapidesa d'aquest mètode no compensa el baix percentatge d'estimacions correctes; el mètode *bagging* aplicat als CART de classificació, tot i estar basat en un classificador simple, clarament hem comprovat que la reducció de la variància per part del mètode d'ensemble fa que la precisió augmenti respecte a l'obtinguda dels arbres (89% a 95%) i finalment, les Màquines de Vectors Suport, a diferència dels altres, pensàvem que ens donaria una estimació elevada i així ho ha fet, al principi estàvem preocupats per si efectuàvem sobre ajustament, però un cop hem usat el conjunt de dades reservat per a la prova d'estimacions finals podem confirmar que és el millor estimador aplicat en aquest conjunt de dades amb una precisió del 97%.

Referències

[1] T. Hastie,R. Tibshirani,J. Friedman,and J. Franklin.(2005) *The Elements of Statistical Learning: data mining, inference and prediction*. Springer.

[2] G. James,D. Witten,T. Hastie and R. Tibshirani.(2013).*An Introduction to Statistical Learning*. 2nd edition.Springer.

[3] Mitchell, T.M.(1997).*Machine Learning*.McGraw Hill.

[4] B. K. Natarajan.(1991).*Machine Learning. A Theoretical Approach*.Morgan Kaufman Publishers.

[5] Hsu,C.-W., Chang,C.-C. and Lin,C.-J.A.(2008).*Practical Guide to Support Vector Classification*.

[6] EmprendePymes.(2013).*Qué es el fraude fiscal, la evasión fiscal y la elusión fiscal?*.[en línia].[consulta: 22 de maig de 2015].Disponible a: `http://goo.gl/QeViXl`

[7] Expansión.com.(2014).*La moral de los españoles, el gran detonante del fraude*.[en línia].[consulta: 5 de juny de 2015]. Disponible a: `http://goo.gl/pZ6mjf`

[8] A. Moore.(2008).*Statistical Data Mining Tutorials*[en línia].[consulta: 20 de juny de 2015].Disponible a: `http://www.cs.cmu.edu/~awm/tutorials`

[9] J. Fortiana.(2014).*Aprenentatge Automàtic i Mineria de Dades*.Universitat de Barcelona.

[10] Universitat de Barcelona.(2015).*Conveni de col·laboració per a la realització de treballs finals de grau d'estudiants de grau de la Facultat de Matemàtiques de la universitat de Barcelona en una empresa o instituacions sense vinculació laboral*.[en línia].Disponible a: `http://goo.gl/t9bwtP`

Annexos

Correlació entre les variables numèriques inicials

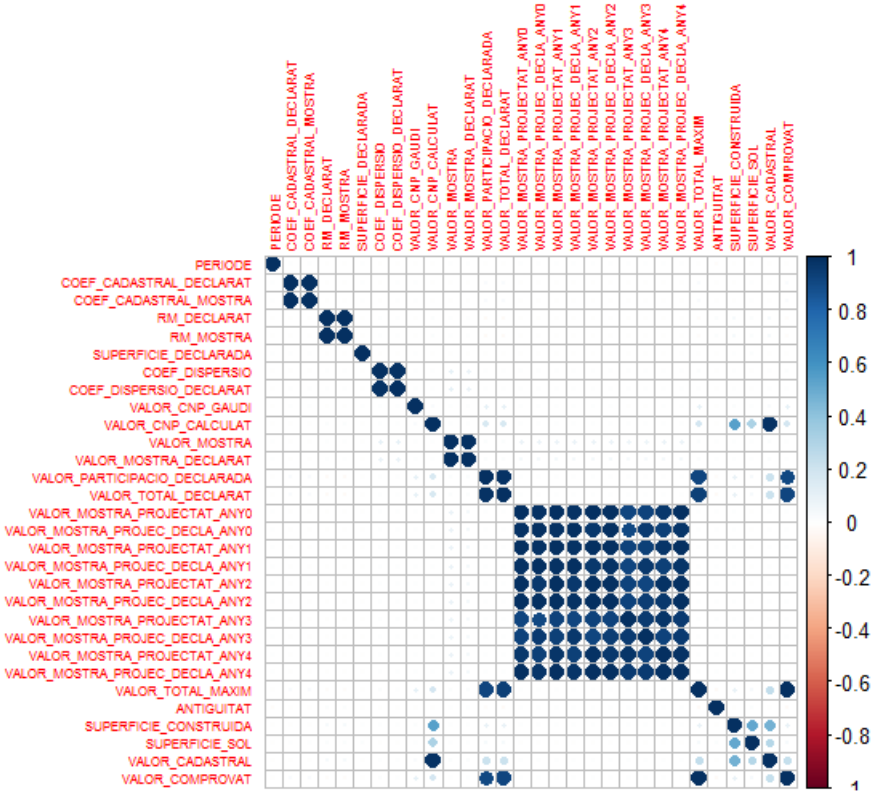

Figura 9: Correlació entre les variables numèriques inicials