Tipologia i cicle de vida de les dades

Pràctica 2

Autors:

Antoni Sànchez Magraner
Joan Solà Porta

1 Descripció del dataset

El dataset triat conté dades relatives a concessions de crèdits bancaris a particulars, concretament 20 atributs. Cada ocurrència és un cas concret de concessió de crèdit, i apareix classificada segons si la persona ha estat bon pagador o, en canvi, ha ocasionat problemes a l'entitat bancària degut a l'impagament del crèdit.

A partir d'aquests tipus de conjunts de dades, dades històriques recollides dels seus propis clients, les entitats bancàries i altres prestadors de crèdit poden avaluar el risc que corren quan concedeixen un crèdit. Aquests tipus d'anàlisis es coneixen com anàlisi del risc de crèdit.

Com es pot observar examinant el dataset, conté atributs categòrics, així com atributs numèrics, tant discrets com continus. A més disposa d'un atribut de classificació en dos grups (booleà), si s'és un bon pagador o ha tingut impagaments. Això permet aplicar-hi tècniques de classificació. A més, obviant aquest atribut de classificació, també es podrien aplicar tècniques de clustering al conjunt de dades (per exemple, si es proven de generar dos clusters, k=2, es podrà comparar amb la classificació ja donada).

El dataset s'ha extret del UCI Machine Learning repository, es tracta d'un dataset elaborat pel *Dr. Hans Hofmann* del *Institut fur Statistik und Okonometrie* de la Universitat de Franfurt.

A continuació es descriuen els camps del conjunt de dades:

- estatCompteCorrent: Estat en què es troba el compte corrent del client. Admet els següents valors:
  - A11 : ... $<$ 0 DM *, on DM són marcs alemanys*
  - A12 : 0 $<=$ ... $<$ 200 DM

  - A13 : ... >= 200 DM
  - A14 : sense compte corrent
- mesosCredit: Temps de retorn del crèdit, en mesos.
- historiaCreditsAnteriors: Històric de crèdits anteriors.
  - A30 : no ha demanat mai cap crèdit, o, tots han estat retornats degudament
  - A31 : els crèdits que va sol·licitar al banc ja han estat retornats degudament
  - A32 : fins ara ha retornat degudament les quotes del crèdit (que encara té actiu)
  - A33 : en algunes ocacions va haver d'endarrerir el pagament del crèdit
  - A34 : no ha retornat alguns pagament, o, té crèdits a altres bancs
- motiu: objecte del crèdit
  - A40 : cotxe nou
  - A41 : cotxe segona mà
  - A42 : mobles
  - A43 : televisor
  - A44 : usos domèstics
  - A45 : reparacions
  - A46 : educació
  - A47 : vacances
  - A48 : reciclatge
  - A49 : negocis
  - A410 : altre
- quantitat (numèric continu): Quantitat de marcs alemanys pels que es sol·licita el crèdit.
- estatCompteEstalvi (categòric): Quantitat de doblers que té al compte corrent
  - A61 : ... < 100 DM
  - A62 : 100 <= ... < 500 DM
  - A63 : 500 <= ... < 1000 DM
  - A64 : .. >= 1000 DM
  - A65 : Sense compte corrent
- tempsTreballActual (categòric): Temps al lloc de feina actual.
  - A71 : descoupat
  - A72 : ... < 1 any
  - A73 : 1 <= ... < 4 anys
  - A74 : 4 <= ... < 7 anys
  - A75 : .. >= 7 anys
- percentRentaDedicat: Percentatge de la renta que es dedica a pagar el deute.

- sexeEstatCivil: Conté informació sobre l'estat civil i el sexe del sol·licitant.
  - A91 : home : divorciat
  - A92 : dona : divorciada o casada
  - A93 : home : solter
  - A94 : home : casat/vidu
  - A95 : dona : soltera
- altresDeutors: indica si el crèdit té algun avalista o altres deudors
  - A101 : no
  - A102 : crèdit conjunt
  - A103 : amb avalista
- tempsResidenciaActual: anys que porta visquen a la residència actual
- propietats: propietats i actius en propietat del sol·licitant
  - A121 : bens immobles
  - A122 : assegurança de vida, pla de pensions (si no té A121)
  - A123 : cotxe (si no té A122 ni A121)
  - A124 : sense cap propietat (o desconeguda)
- edat: edat, en anys, del sol·licitant
- altresPlans: propietats i actius en propietat del sol·licitant
  - A141 : bank
  - A142 : stores
  - A143 : none
- propietatVivendaActual: propietat de la vivenda principal
  - A151 : llogada
  - A152 : pròpia
  - A153 : hi viu gratuïtament
- numCredits: quantitat de crèdits que ja atorgats la persona.
- tipusFeina: tipus d'ocupació de la persona.
  - A171 : a l'atur
  - A172 : professió no qualificada
  - A173 : professió qualificada
  - A174 : gestor o professió altament qualificada
- numPersonesManteniment: número de persones que intervendran en el crèdit
- teTelèfon: si la persona disposa de telèfon enregistrat al seu nom
  - A191 : no

  - A192 : sí

- estranger: si la persona és alemanya o estrangera
  - A191 : no
  - A192 : sí

- bonPagador: element de classificació si és o no bon pagador:
  - 1 : sí
  - 2 : no

Com s'ha comentat aquests tipus de conjunts de dades són altament valuosos per les entitats bancàries ja que permeten crear models per predir les probabilitats d'impagament d'un crèdit i així decidir si l'atorgament dels crèdits. Per tant, el problema a resoldre a través d'aquests conjunts de dades és determinar els valors dels atributs que millor defineixen un bon pagador i els que defineixen els mals pagadors.

2 Integració i selecció de dades

El conjunt de dades seleccionat conté un conjunt suficient i adequat d'atributs per poder realitzar anàlisis a partir dels quals poder determinat les característiques comuns que solen tenir els clients que generen problemes de pagaments i crear un model predictiu per determinar la probabilitat d'impagament depenent dels determinats atributs.

Tot i això, si el conjunt de dades incorporàs un camp que identifiqués de manera únivoca als clients (per exemple un DNI), seria possible completar el conjunt de dades amb altres fonts d'informació sobre els respectius clients (per exemple un conjunt de dades de l'activitat dels clients en el portal web del banc, o de l'ús que en fan de les targetes de crèdit i dèbit). Per tant, el fet que el conjunt de dades no incorpori cap camp que identifiqui al client no permet complementar-ho a nivell de cada registre.

Per això, l'anàlisi s'ha realitzat sense haver descartat, d'entrada, cap dels atributs que conté. Però tampoc no s'han incorporat dades o camps nous al conjunt inicial. El que sí s'ha desenvolupat és l'anàlisi de la qualitat de les dades i la corresponent neteja, tal i com es detallarà al següent apartat.

3 Neteja de dades

3.1 Anàlisi preliminar del conjunt de dades

La primera passa abans de poder realitzar qualsevol procés d'anàlisi i mineria sobre el conjunt de dades és analitzar la seva qualitat i fer-ne les correccions i ajustaments necessaris, per tal que els resultats de les anàlisis de dades posteriors tinguin la màxima qualitat i no estiguin esbiaixats.

En primer lloc s'examina de manera global el conjunt de dades per disposar d'informació bàsica de cadascun dels camps, aplicant la funció summary, que descriu tots els camps:

```
##  estatCompteCorrent  mesosCredit   historiaCreditsAnteriors     motiu
##  A11:274            Min.   : 4.0   A30: 40                  A43    :280
##  A12:269            1st Qu.:12.0   A31: 49                  A40    :234
##  A13: 63            Median :18.0   A32:530                  A42    :181
##  A14:394            Mean   :20.9   A33: 88                  A41    :103
##                     3rd Qu.:24.0   A34:293                  A49    : 97
##                     Max.   :72.0                            A46    : 50
##                                                             (Other): 55
##    quantitat     estatCompteEstalvi tempsTreballActual percentRentaDedicat
##  Min.   :  250   A61:603            A71: 62            Min.   :1.000
##  1st Qu.: 1366   A62:103            A72:172            1st Qu.:2.000
##  Median : 2320   A63: 63            A73:339            Median :3.000
##  Mean   : 3271   A64: 48            A74:174            Mean   :2.973
##  3rd Qu.: 3972   A65:183            A75:253            3rd Qu.:4.000
##  Max.   :18424                                         Max.   :4.000
##
##  sexeEstatCivil altresDeutors tempsResidenciaActual propietats
##  A91: 50        A101:907      Min.   :1.000         A121:282
##  A92:310        A102: 41      1st Qu.:2.000         A122:232
##  A93:548        A103: 52      Median :3.000         A123:332
##  A94: 92                      Mean   :2.845         A124:154
##                               3rd Qu.:4.000
##                               Max.   :4.000
##
##       edat       altresPlans propietatVivendaActual   numCredits
##  Min.   :19.00   A141:139    A151:179               Min.   :1.000
##  1st Qu.:27.00   A142: 47    A152:713               1st Qu.:1.000
##  Median :33.00   A143:814    A153:108               Median :1.000
##  Mean   :35.55                                      Mean   :1.407
##  3rd Qu.:42.00                                      3rd Qu.:2.000
##  Max.   :75.00                                      Max.   :4.000
##
##  tipusFeina numPersonesManteniment teTelèfon  estranger    bonPagador
##  A171: 22   Min.   :1.000          A191:596   A201:963   Min.   :1.0
##  A172:200   1st Qu.:1.000          A192:404   A202: 37   1st Qu.:1.0
##  A173:630   Median :1.000                                Median :1.0
##  A174:148   Mean   :1.155                                Mean   :1.3
##             3rd Qu.:1.000                                3rd Qu.:2.0
##             Max.   :2.000                                Max.   :2.0
##
```

Figura 1: Resultat de la funció summary sobre el conjunt de dades.

A més es calcula el nombre de valors diferents de cada variable:

```
##      estatCompteCorrent             mesosCredit historiaCreditsAnteriors
##                       4                      33                       5
##                   motiu                quantitat      estatCompteEstalvi
##                      10                     921                       5
##      tempsTreballActual     percentRentaDedicat          sexeEstatCivil
##                       5                       4                       4
##           altresDeutors    tempsResidenciaActual              propietats
##                       3                       4                       4
##                   edat             altresPlans   propietatVivendaActual
##                      53                       3                       3
##             numCredits              tipusFeina   numPersonesManteniment
##                       4                       4                       2
##              teTelèfon               estranger              bonPagador
##                       2                       2                       2
```

Figura 2: Diversitat de valors de cada variable.

Observant el resum de camps, es pot veure que:

- El camp bonPagador està definit com a numèric: (1 - bon pagador i 2 - mal pagador). Aquesta columna es transformarà en termes de 0 i 1, per convertir-la a factor booleà.

A més, aplicant el summary a la variable s'observa que al voltant d'un 70% dels registres es corresponent a bons pagadors.

- Com és lògic, els atributs que admeten una quantitat major de valors diferents són numèrics: quantitat (sol·licitada), edat, mesosCredit (durada en mesos). Però alguns atributs cosiderats numèrics a la font de dades original, podríen ser considerats com factors, degut al seu reduït nombre de valors possibles. Són concretament els següents:
  - percentRentaDedicat: fins a quatre.
  - tempsResidenciaActual: fins a quatre.
  - numCredits: fins a quatre crèdits.
  - numPersonesManteniment: 1 (el propi titular) o 2 persones. Per tant, en aquest cas es transforma en 0 o 1, depenent de sí hi ha un segon titular (1) o no (0).
- Observant els valors d'edat, quantitat i mesosCredit, es pot intuir que aquestes variables podrien contenir valors que surtin de la normalitat. Per exemple, en el cas de la quantiat: que la seva mitjana sigui 3.271, la mediana 2.320 i el tercer quartil 3.972, sembla que el valor màxim 18.424, s'allunya bastant d'aquests valors. Per tant, s'haurà d'estudiar més en profundidat si s'han d'eliminar algunes ocurrències (detecció de outliers). A continuació s'analitzaran els outliers, i s'estudiarà la conveniència d'eliminar els registres que els contenguin.
- Els camps categòrics estan codificats seguint una notació A0..90..90..9, per tant es realitza la conversió d'aquests valors codificats a valors semàntics, amb l'objectiu de facilitar la seva comprensió.
- Addicionalment es cerquen valors nuls, zeros o buits al conjunt de dades. Com es pot observar, el resultat d'aquesta certca és negatiu, el conjunt de dades no en conté cap. Per tant, es farà cap correcció al respecte.

```
> colSums(is.na(credit_ds) )
      estatCompteCorrent             mesosCredit historiaCreditsAnteriors                   motiu
                       0                       0                       0                       0
               quantitat      estatCompteEstalvi       tempsTreballActual      percentRentaDedicat
                       0                       0                       0                       0
          sexeEstatCivil           altresDeutors    tempsResidenciaActual              propietats
                       0                       0                       0                       0
                    edat             altresPlans   propietatVivendaActual             numCredits
                       0                       0                       0                       0
              tipusFeina   numPersonesManteniment                telefon               estranger
                       0                       0                       0                       0
             bonPagador
                       0
> colSums(credit_ds=="")
      estatCompteCorrent             mesosCredit historiaCreditsAnteriors                   motiu
                       0                       0                       0                       0
               quantitat      estatCompteEstalvi       tempsTreballActual      percentRentaDedicat
                       0                       0                       0                       0
          sexeEstatCivil           altresDeutors    tempsResidenciaActual              propietats
                       0                       0                       0                       0
                    edat             altresPlans   propietatVivendaActual             numCredits
                       0                       0                       0                       0
              tipusFeina   numPersonesManteniment                telefon               estranger
                       0                       0                       0                       0
             bonPagador
                       0
```

Figura 3: Resultat de la cerca de valors nuls o buits.

3.2 Anàlisi d'outliers

Com s'ha comentat a l'anàlisi preliminar de la qualitat de dades, a continuació es realitzarà un estudi dels possibles outliers, centrats en les variables per les que s'han tingut indicis

de valors anormals a l'anàlisi preliminar, per això s'empraran diverses tècniques, tant gràfiques com tests.

Outliers de *quantitat*

- Anàlisi d'outliers de la variable *quantitat* mitjançant gràfics de caixes: Com es pot observar al gràfic, tot i que la majoria de valors es concentren entre el rang 1.000 - 5.000DM, alguns valors estan per sobre dels 8.000, i podrien ser considerats outliers.

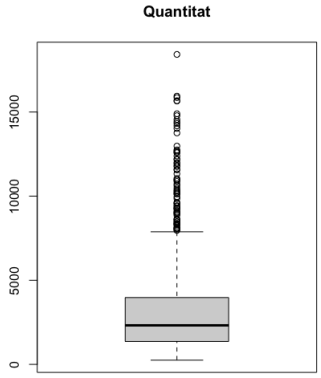

Figura 4: Gràfic de caixa de la variable quantitat.

- Detecció d'outliers a través del criteri de les dues desviacions estàndard: s'han marcat els registres en què el valor de la variable *quantitat* es troba, unidimensionalment, a més de 2 desviacions estàndard de la mitjana. Segons aquesta tècnica, els valors per damunt de 8.947 podran ser considerats outliers (en total es detecten 55 outliers de *quantitat*).

Outliers d'*edat*

- Anàlisi d'outliers de la variable *edat* mitjançant gràfics de caixes: Com es pot observar al gràfic, tot i que la majoria de valors es concentren entre el rang 25 - 45, alguns valors estan per sobre dels 60, que podrien ser considerats outliers.

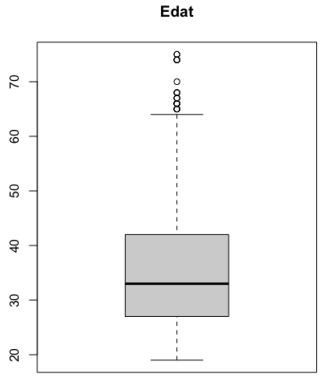

Figura 5: Gràfic de caixa de la variable edat.

- Detecció d'outliers a través del criteri de les dues desviacions estàndard: s'han marcat els registres en què el valor de la variable *edat* es troba, unidimensionalment, a més de 2 desviacions estàndard de la mitjana. Segons aquesta tècnica, les edats iguals o superiors a 59 anys podrien ser considerades outliers (en total es detecten 54 outliers d'*edat*).

Outliers de *mesosCredit*

- Anàlisi d'outliers de la variable *mesosCredit* mitjançant gràfics de caixes: Com es pot observar al gràfic, tot i que la majoria de valors es concentren entre el rang 10 - 30, alguns valors estan per sobre dels 40, que podrien ser considerats outliers.

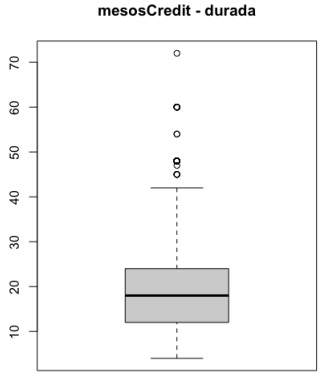

Figura 6: Gràfic de caixa de la variable edat.

- Detecció d'outliers a través del criteri de les dues desviacions estàndard: s'han marcat els registres en què el valor de la variable *mesosCredit* es troba, unidimensionalment, a més de 2 desviacions estàndard de la mitjana. Segons aquesta tècnica, els temps de retorn del crèdit iguals o superiors a 47 mesos podrien ser considerades outliers (en total es detecten 65 outliers de *mesosCredit*).

Tenint en compte que els valors detectats com a outliers representen valors extrems i representen menys d'un 5% del conjunt total de dades de cada variable, es pot considerar que aquests valors podrien desvirtuar i dificultar les anàlisis del conjunt de dades en general, per tant, es decideix crear un nou conjunt de dades a partir de l'eliminació dels registres que contenen valors d'alguna variable detectats com a outliers, el conjunt de dades resultant conté un total de 853 registres, per tant, es descarten 147 registres del conjunt de dades original (alguns registres contenien valors detectats com a outliers en més d'una de les variables estudiades).

A més de les tècniques emprades per a determinar els outliers a diferents variables, de manera individual, també es pot emprar la distància de Mahalanobis per determinar la similitud entre variables aleatòries multidimensionals. Per exemple, s'aplicarà la distància de Mahalanobis per detectar possibles outliers del conjunt bidimensional de dades format per les variables *quantitat* i *edat*.

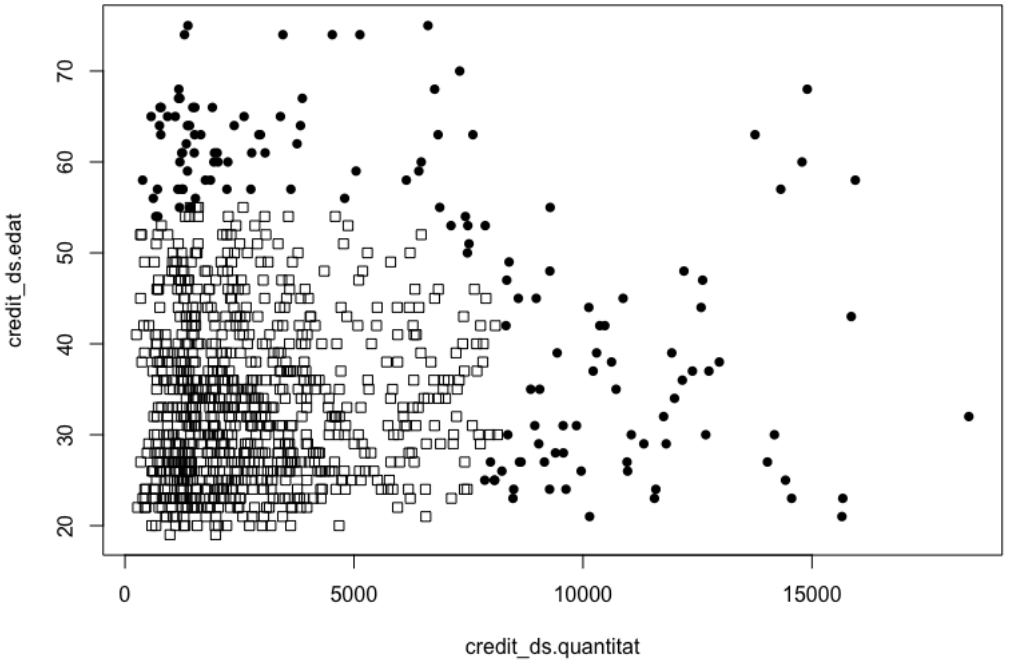

Figura 7: Representació gràfica dels valors de les variables edat i quantitat, amb els primers 147 valors més distants de la resta marcats com a punts negres.

3.3 Normalització de variables numèriques

Depenent de la diferència entre magnituds dels diferents camps pot ser convenient escalarlos per homogenaïtzar-los. Per tant, addicionalment s'afegeixen al conjunt de dades nous camps creats a partir d'aplicar la tècnica de normalització (valors entre (-1 i 1) sobre cadascun dels conjunts de dades. En definitiva, es tracta de reduir el biaix causat per la combinació de valors mesurats a diferents escales.

4 Anàlisi de les dades

4.1 Selecció i anàlisi de grups de dades

Una pregunta que sorgeix de manera natural del conjunt de dades carregat és si hi ha diferències entre els bons i mals pagadors de crèdits. En aquest projecte, doncs, ens proposem, a partir de tests estadístics i visualitzacions, entendre les possibles diferències entre aquests dos grups de persones.

En les següents subseccions ens centrem en la part d'aplicació de les tècniques estadístiques.

4.2 Anàlisi gràfic de les dades

A continuació s'analitzarà a través de gràfiques la relació entre la variable bon pagador i alguns dels atributs més significatius, tant categòrics com numèrics.

4.2.1 Relació amb l'edat

De les següents gràfiques, es visualitza clarament que la majoria de credits es sol·liciten entorn als 30 anys (entre els 25 i 35 anys). A més, el percentatge d'impagaments (mal pagadors) en general és lleugerament més baix com més edat tenen els clients.

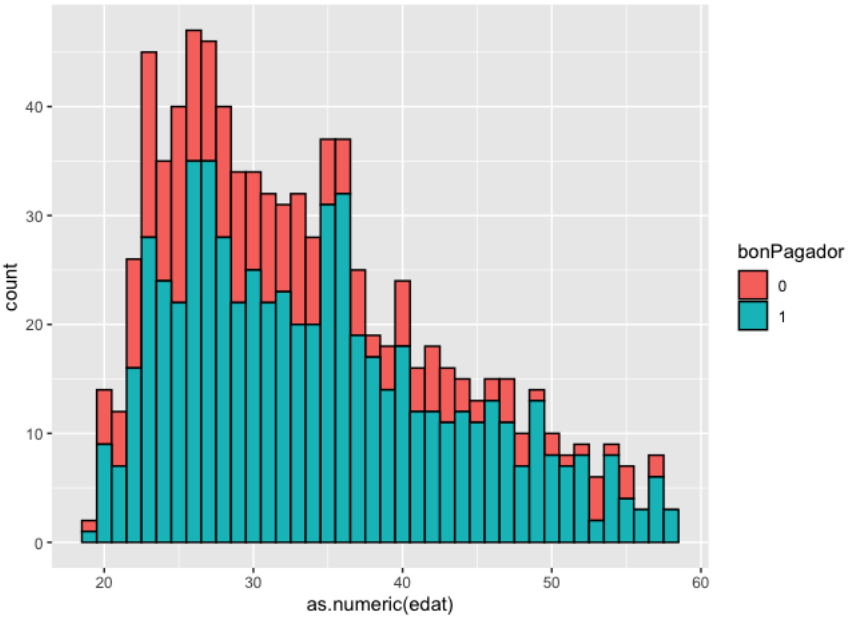

Figura 8: Relació quantitativa entre edat i la morositat.

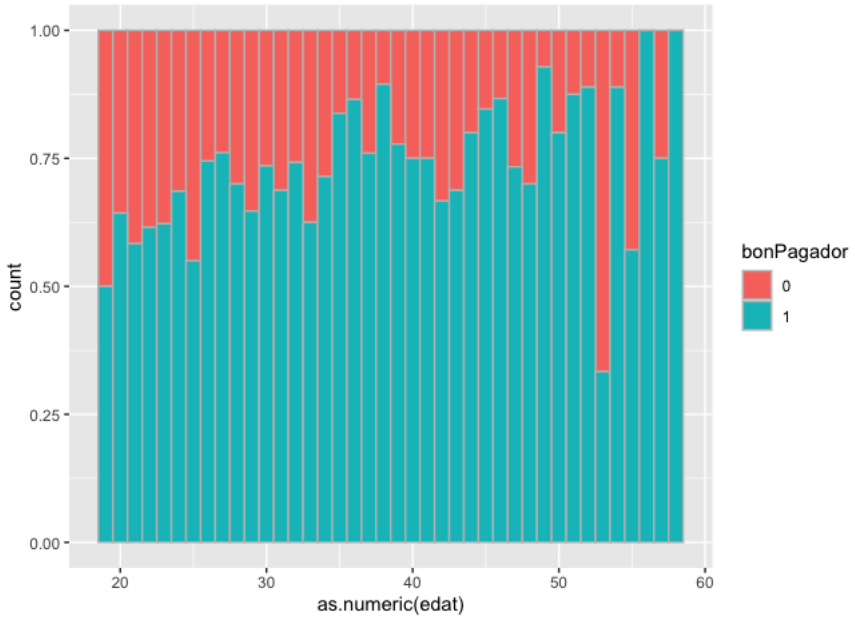

Figura 9: Relació relativa entre edat i la morositat.

4.2.2 Relació amb la quantitat de crèdit

Les gràfiques indiquen que no hi ha una relació directa o concloent entre la quantitat sol·licitada i la taxa d'impagaments (per a crèdits molt petits o molt alts la taxa d'impagaments sembla ser lleugerament superior als crèdits d'entre 1.500 i 4.000 €)

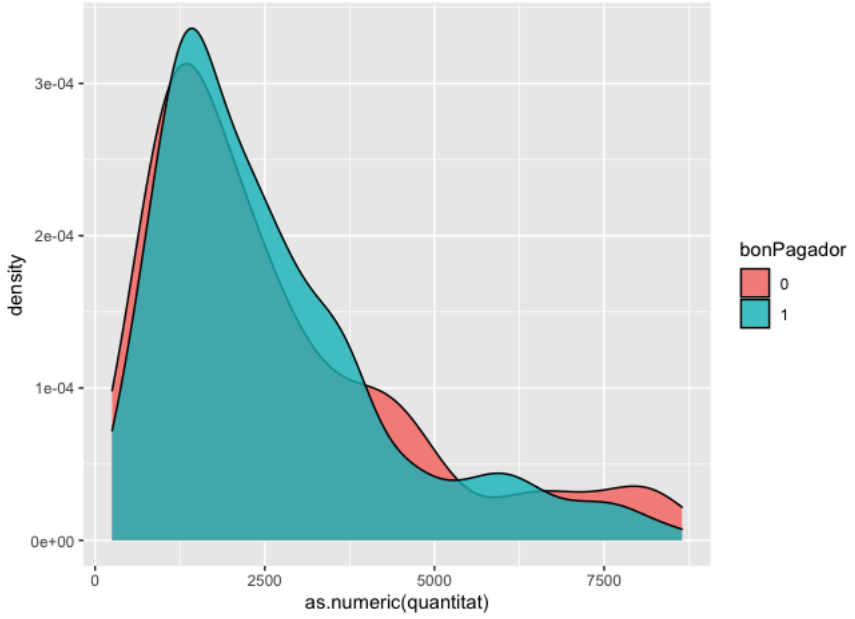

Figura 10: Relació quantitativa entre quantitat de crèdit i la morositat.

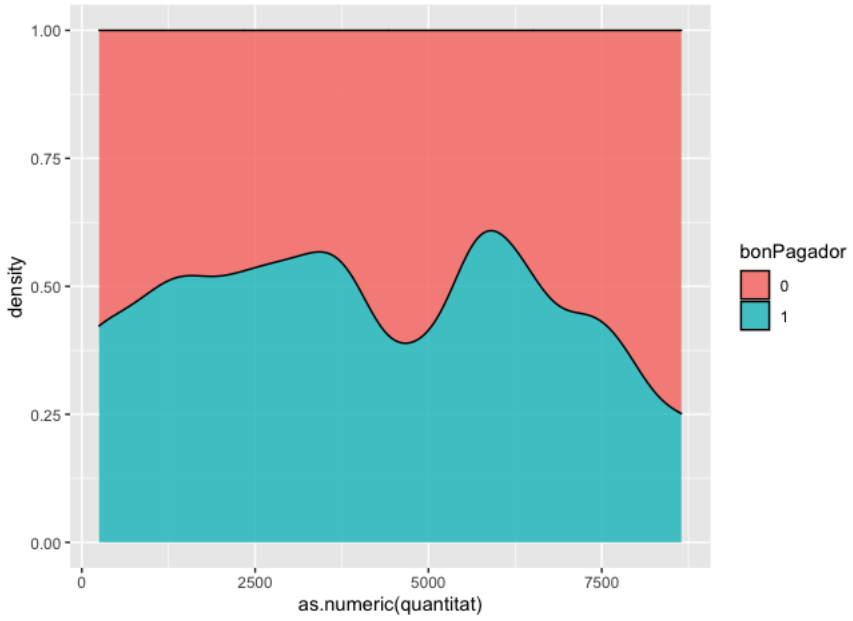

Figura 11: Relació relativa entre quantitat de crèdit i la morositat.

Aquesta conclusió es pot veure més clarament les gràfiques es generen a partir d'una nova nova variable derivada de la discretització per k-means de la variable quantitat.

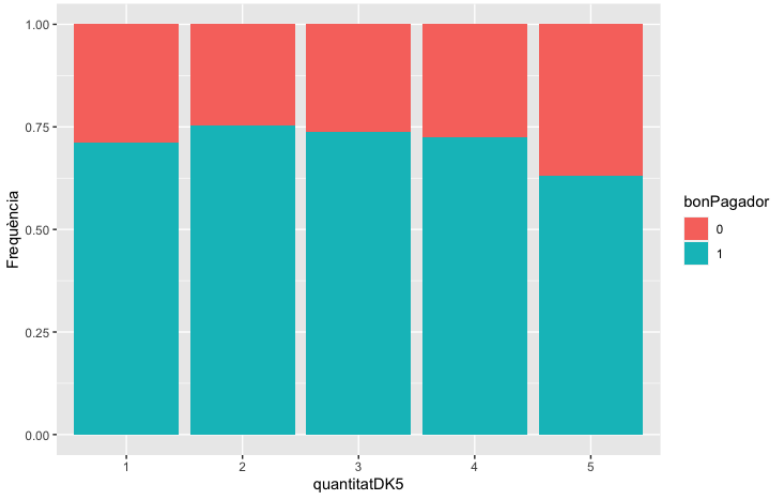

Figura 12: Relació relativa entre quantitat discretitzada i la morositat.

4.3 Comprovació de la normalitat i homogeneïtat de la variància.

L'aplicació d'algunes tècniques com les proves per contrast d'hipòtesis de tipus paramètric, depenen de la distribució de la variable i de la seva homoscedasticitat. Per tant, a continuació es comprovarà en primer lloc si les variables numèriques segueixen una distribució normal i a continuació la seva homoscedasticitat.

4.3.1 Comprovació de la normalitat

Per a comprovar la normalitat s'empraran els tests de Kolmogorov-Smirnov i Shapiro-Wilk, a més de la tècnica de representació gràfica Q-Q.

Comprovació de la normalitat de *quantitat*

- L'aplicació d'ambdós tests de normalitat sobre la variable *quantitat* amb els seus outliers ja eliminats retornen valors p molt menors als nivell de significació, per tant la hipòtesi nul·la seria rebutjada i, per tant, es conclou que quantitat no segueix una distribució normal.

```
> shapiro.test(credit_ds_net$quantitat)

        Shapiro-Wilk normality test

data:  credit_ds_net$quantitat
W = 0.87225, p-value < 2.2e-16
```

Figura 13: Aplicació del test Shapiro-Wilk a la variable quantitat.

- A través de la valoració gràfica emprant Q-Q plot, es pot comprovar que els punts no segueixen la forma de la línia recta, per tant, a través d'aquesta tècnica també es confirma que no es tracta d'una distribució normal.

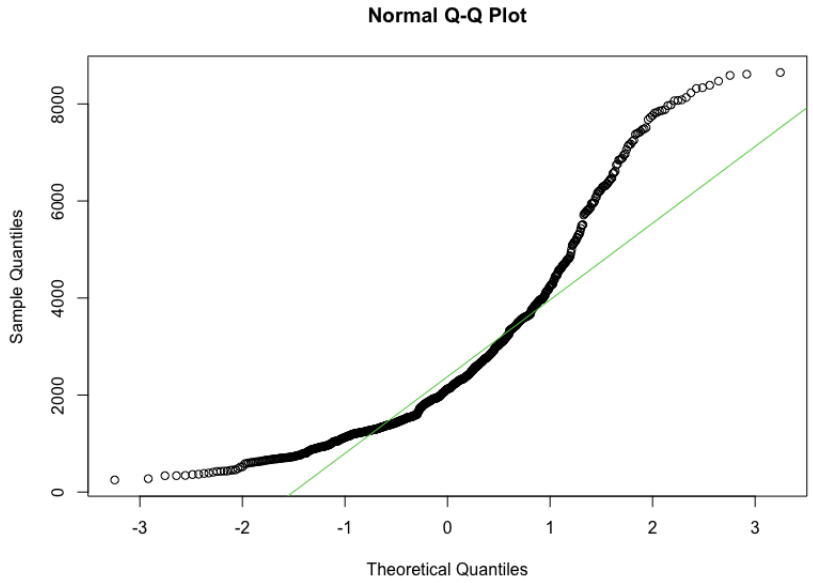

Figura 14: Q-Q Plot de la variable quantitat.

Comprovació de la normalitat d'*edat*

- L'aplicació d'ambdós tests de normalitat sobre la variable *edat* amb els seus outliers ja eliminats retornen valors p molt menors als nivell de significació, per tant la hipòtesi nul·la també seria rebutjada en aquest cas i, per tant, es conclou que *edat* tampoc no segueix una distribució normal.

- A través de la valoració gràfica emprant Q-Q plot, es pot comprovar que els punts no segueixen la forma de la línia recta, per tant, a través d'aquesta tècnica també es confirma que no es tracta d'una distribució normal.

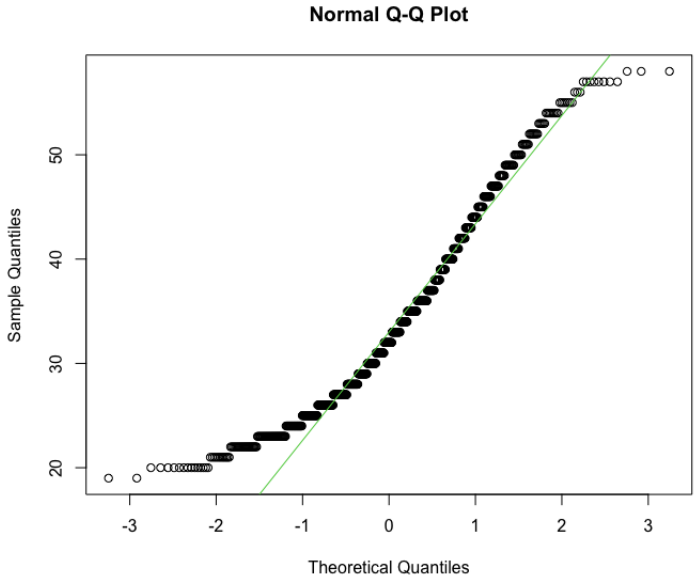

Figura 15: Q-Q Plot de la variable edat, havent eliminat els outliers

- Addicionalment també s'ha aplicat la tècnica gràfica sobre la variable no netejada (incloent els valors detectats com a outliers), en aquest cas es pot veure de manera gràfica que l'anterior gràfica (sense valors outliers), s'aproxima més a la línia recta, que l'obtinguda incloent els outliers.

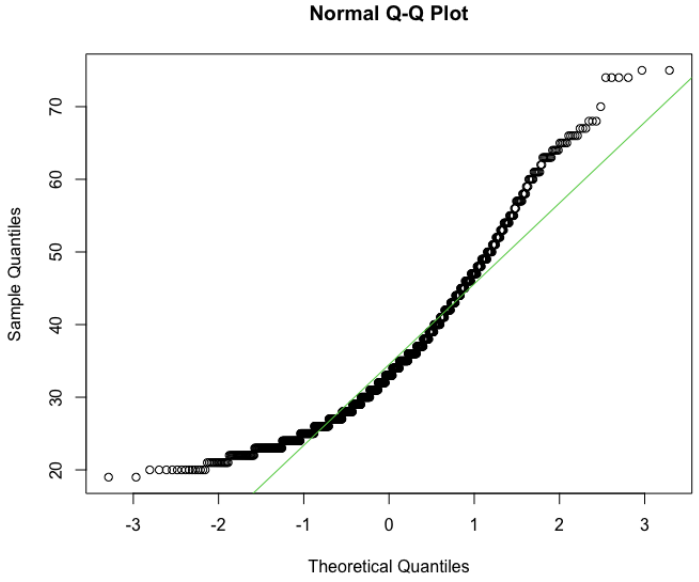

Figura 16: Q-Q Plot de la variable edat, incloent outliers.

Comprovació de la normalitat de *mesosCredit*

- L'aplicació d'ambdós tests de normalitat sobre la variable *mesosCredit* amb els seus outliers ja eliminats retornen valors p molt menors als nivell de significació, per tant la hipòtesi nul·la també seria rebutjada en aquest cas i, per tant, es conclou que *mesosCredit* tampoc no segueix una distribució normal.
- A través de la valoració gràfica emprant Q-Q plot, es pot comprovar que els punts no segueixen la forma de la línia recta, per tant, a través d'aquesta tècnica també es confirma que no es tracta d'una distribució normal.

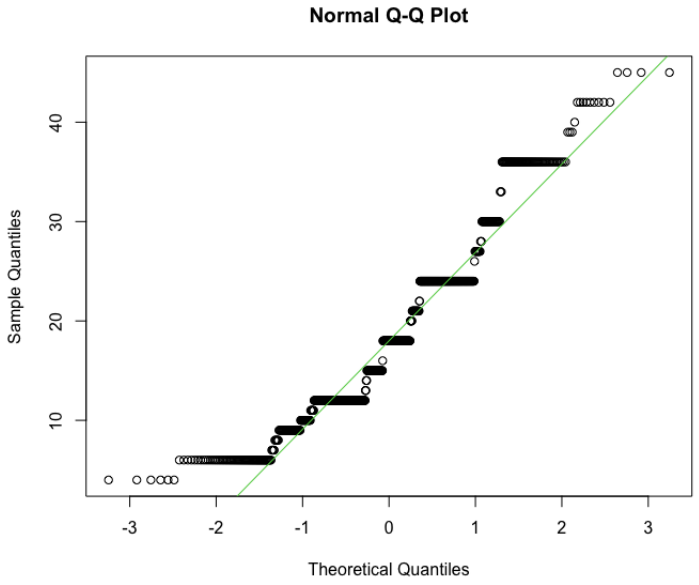

Figura 17: Q-Q Plot de la variable mesosCredit

- En aquest cas, observant la gràfica es pot veure que té un aspecte escalonat, degut a què hi ha menys valors diferents en el conjunt de mesos (en comparació amb altres variables com quantitat).

4.3.2 Comprovació de la homoscedasticitat

En aquest apartat s'ha procedit a fer el test F d'homogeneïtat de la variància[1] entre el grup de bons i mals pagadors. En aquest test es tenen dues hipòtesis:

- Hipòtesi nul·la (H0): les variàncies dels dos grups són iguals.
- Hipòtesi alternativa (H1): les variàncies dels dos grups són diferents.

A continuació s'inclu una taula amb els p-valors i la hipòtesi amb que ens quedem en cada cas (cal notar que només descartarem H0 en cas de que p-valor$<$0.05):

[1]Veure la referència [1] per a més informació

Variable	p-valor	Hipòtesi
quantitat	0.01564	H1
edat	0.6184	H0
Mesos de crèdit	0.2622	H0
Percentatge de la renda dedicat	0.2594	H0
Temps en la residència actual	0.9132	H0
Número de crèdits	0.9132	H0

Així doncs, a partir dels tests només es pot afirmar que existeix una diferència significativa entre les variàncies de la variable quantitat en comparar els grups de bons i mals pagadors.

4.4 Proves estadístiques de comparació

4.4.1 Proporció de clients que tenen un immoble en propietat

Una intuïció que es pot tenir abans d'examinar les dades és que aquelles persones que tinguin alguns propietat immobiliària seran aquelles amb més probabilitat de ser bons pagadors. Justament per tal de comprovar si aquesta suposició és certa, es planteja el següent test d'hipòtesis unilateral:

- Hipòtesi nul·la (H0): la proporció de persones amb propietats immobiliàries en el grup de bons pagadors és igual o menor a la del grup de mals pagadors.
- Hipòtesi alternativa (H1): la proporció de persones amb propietats immobiliàries en el grup de bons pagadors és superior a la del grup de mals pagadors.

Tal com es pot observar si s'executa el codi de la pràctica, obtenim que el p-valor d'aquest test és $4.825374 \cdot 10^{-3}$, de manera que es por descartar la hipòtesi nul·la i acceptar l'alternativa. En conclusió, doncs, es pot afirmar que els bons pagadors tenen més probabilitat de tenir propietats immobiliàries que els mals pagadors

4.4.2 Intervals de confiança de la mitjana de la quantitat demananda

Tal com s'ha pogut veure a l'apartat 4.2, la variable 'quanitat' no segueix una distribució normal, la qual cosa no ens permet fer el test de comparació de mitjanes d'aquesta variable per a bons i mals pagadors. Això és degut a que aquesta és una condició que s'ha de complir per poder aplicar el test[2].

Tot i això, si ens basem en el teorema del límit central, el que sí que podem fer és buscar els intervals de confiança del 95% per a la mitjana d'aquesta variable en ambdós grups i veure si se solapen. Amb aquest procediment, els resultats obtinguts són els següents:

Grup	Límit inferior (95% conf.)	Límit superior (95% conf.)
Bons pagadors	2470.252	2744.475
Mals pagadors	2445.938	2954.609

Veient els resultats, no sembla que hi hagi cap indicació de que els dos grups comparats tinguin mitjanes diferents pel que fa a la quantitat dels crèdits demanats.

[2]Veure més informació a la referència [1].

4.4.3 Correlació de variables

Una altra pregunta que val la pena fer-nos és quines de les variables estan correlades o no amb ser bon o mal pagador. Per intentar respondre aquesta pregunta, ens cal diferenciar entre variables contínues i categòriques, ja que els tests que podrem fer sobre unes i altres són diferents.

En el cas de les variables numèriques, podem fer el test de correlació d'Spearman entre aquestes i la variable que defineix els grups. Els resultats obtinguts són els següents:

Variable	p-valor	coef. correlació
Edat	0.0002533	0.1249697
Mesos de crèdit	$4.495 \cdot 10^{-06}$	-0.1563218
Quantitat	0.7244	0.01208979

Com es pot veure, a partir dels resultats es pot afirmar que hi ha una correlació no nula entre l'edat dels clients i el temps de durada del crèdit i ser o no un bon pagador.

Per al cas de la correlació entre el ser bon pagador i les variables categòries s'han emprat els tests tetrachoric (per a variables binàries, que només admeten dos valors) i Cramer V (per a la resta de variables). Per aplicar aquests tests en primer lloc es crea una matriu de relació entre els valors possibles d'ambdues variables, per passar-la com a paràmetre a la respectiva funció, per exemple per al cas de la variable *propietats*:

```
> matriu_propietatsXbonPagador <- with(credit_ds_net, table(propietats, bonPagador))
> matriu_propietatsXbonPagador
                                                bonPagador
propietats                                        0   1
  building society savings agreement/ life insurance  63 152
  car or other                                      80 206
  real estate                                       54 199
  unknown/no property                               37  62
```

Figura 18: Taula de relació entre els valors de les variables propietats i bonPagador

A continuació es mostren els valors obtinguts de l'aplicació dels tests esmentats sobre les variables categòriques més rellevants:

Variable	test	coef. correlació
Propietats	Cramer V	0.1085
Immoble	Cramer V	0.08861
Immoble	Tetrachoric	0.16
Estat compte corrent	Cramer V	0.3423
Estat compte estalvi	Cramer V	0.1659
Històric crèdits anteriors	Cramer V	0.247
Motiu del crèdit	Cramer V	0.2177
Temps feina actual	Cramer V	0.1707
Sexe i estat civil	Cramer V	0.1144
Tipus de feina	Cramer V	0.03811

Com es pot observar a partir dels diferents coeficients de correlació obtinguts de cada variable, en general les correlacions són bastant dèbils, destaca especialment la correlació

entre l'estat del compte corrent i el ser o no bon pagador.

4.4.4 Regressió logística

Vistes les diferències detectades fins ara entre els conjunts de bons i mals clients i les correlacions de les diverses variables del conjunt de dades amb el fet d'estar en un dels dos grups, ens preguntem si seriem capaços de predir quines persones seran bones o males pagadores. Per tal d'aconseguir-ho, crearem un model de regressió logística que tindrà com a variable objectiu ser bon o mal pagador.

Un cop entrenat el model, veiem que la seva matriu de confusió és la següent:

		Valors predits	
		Mal pagador	Bon pagador
Valors correctes	Mal pagador	117	117
	Bon pagador	51	568

A partir d'aquí, doncs, s'obtenen les següents mètriques de rendiment del model:

Mètrica	Valor
Sensibilitat	91.76%
Especificitat	50%

A la taula presentada a continuació es resumeixen els p-valors del test de significància, que ens indica si les variables emprades en el model tenen o no un pes significatiu (direm que una variable és significativa si el p-valor del test és inferior a 0.05), per a les variables que són significatives.

Mètrica	p-valor
mesos crèdit	0.0.000333
és estranger	0.017627
no té altres deutes	0.020494
aporta un avalador	0.004070
és home solter	0.026666
% renda dedicat al pagament	0.009539
persona a l'atur	0.043154
sense comte corrent a l'entitat	5.02e-06
motiu crèdit: compra cotxe nou	0.007413
motiu crèdit: educació	0.009643
crèdits anteriors: en estat crític amb l'entitat/altres crèdits en una altra entitat	0.000396
crèdits anteriors: tots pagats diligentment	0.015658

5 Resolució del problema

Tal com s'havia comentat al principi del document, l'objectiu d'aquest projecte era comparar les característiques de bons i mals pagadors i veure si era possible predir amb certa confiança si un client pertany a un dels dos grups.

A partir d'una comparativa qualitativa dels dos grups, hem pogut veure que:

- Els mals pagadors tendeixen a ser clients més joves.
- No s'observa una relació clara entre impagaments i la quantitat demanada en el crèdit.
- Sembla que les persones amb una propietat immobiliària tenen una menor probabilitat d'impagar.

Amb els tests estadístics emprats, s'ha pogut comprovar que:

- Les persones que paguen els seus crèdits sense problemes tenen una probabilitat major de tenir propietats immobiliàries.
- No hi ha cap indicació clara que les mitjanes de les quantitats demanades per bons i mals pagadors siguin diferents.
- Hi ha una correlació estadísticament significativa entre l'edat del client i els mesos de durada del crèdit amb ser bon o mal pagador.

Com a part final del nostre informe, hem aplicat un model de regressió logística per tal de predir bons i mals clients. Amb aquest s'han obtingut uns resultats prou acceptables (sensibilitat del 91.76% i especificitat del 50%).

Així doncs, es pot concloure que els resultats obtinguts al llarg de la pràctica ens han permès resoldre el problema que ens plantejàvem. Això ho podem dir perquè hem sigut capaços de detectar diferències significatives entre els dos grups analitzats i de fer-les servir per fer prediccions.

6 Codi

Enllaç al repositori GitHub: `https://github.com/jsolatsanchez/Practica2AnalisiDataset`.

7 Dataset

Enllaç de l'adreça del dataset: `https://archive.ics.uci.edu/ml/machine-learning-databases/statlog/german/german.data`.

Referències

[1] Bernadó, Ester. n.d. *"Tests d'hipòtesis"*[en línia]. Barcelona: UOC (S.d.).

Contribucions al treball

Contribucions	Signatura
Investigació prèvia	A.S.M, J.S.P
Redacció de les respostes	A.S.M, J.S.P
Desenvolupament del codi	A.S.M, J.S.P