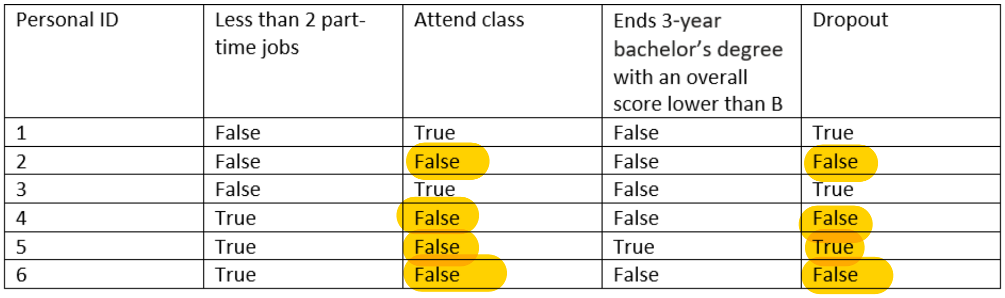
ECO421
University of Toronto

Marlène Koffi

October 7, 2021

1 Exercise 1 (15 points): Construct a decision tree



Personal ID	Less than 2 part-time jobs	Attend class	Ends 3-year bachelor's degree with an overall score lower than B	Dropout
1	False	True	False	True
2	False	False	False	False
3	False	True	False	True
4	True	False	False	False
5	True	False	True	True
6	True	False	False	False

The table shows a fictitious dataset that describes the dropout rate for students at a 4-year Bachelor's degree.

1. (12 points) Using the table and an entropy-based information gain, construct a decision tree (by hand, i.e make the calculus and find the relevant splits) that would predict the dropout status for a student. (NB: the logarithm to use in the entropy measurement is the logarithm to the base 2.)

2. (1.5 points) What will be the prediction generated by the tree for: "Less than 2 part time job"= false, "Attend class"= False and "Ends 3rd degree with an overall score lower than B"=true.

3. (1.5 points) What will be the prediction generated by the tree for: "Less than 2 part time job"= True, "Attend class"= True and "Ends 3rd degree with an overall score lower than B"=true.

2 Exercise 2 (35 points): Logistic versus Linear Discriminant analysis

2.1 Part 1

Consider a two-class logistic regression problem with only one predictor $x \in \mathbb{R}$. We assume that the sample for the two classes are separated by a point 0:

$$Y = \begin{cases} 1 \text{ if } x > 0 \\ \\ 0 \text{ if } x < 0 \end{cases}$$

1. (3 points) Write the likelihood function.

2. (7 points) Assume the data are centered ($\beta_0 = 0$). Characterize the solution estimates of the parameter in this case. Comment.(Hint: Explain why the maximum likelihood estimation does not converge in this case using the expression of the likelihood and the fact that:

   $$\lim_{\beta \to \infty} p(x) = \begin{cases} 1 \text{ if } x > 0 \\ \\ 0 \text{ if } x < 0 \end{cases}$$

   ;

   NB: I abstract for subscript i for simplicity.)

2.2 Part 2

The previous situation illustrates the problem of logistic regression when classes are well separated. Because of this limitation (among other things), we often prefer using instead another classifier called the Linear Discriminant Analysis. Let us understand the linear discriminant analysis.

Consider a two-class logistic regression problem with only one predictor $x \in \mathbb{R}$. The idea of the Linear Discriminant Analysis is to predict the probability $P(Y = k|X = x)$ by assuming that $P(X = x|Y = k)$, i.e. the density function of X conditional on $Y = k$, follows a normal distribution with mean μ_k and variance σ_k^2.

1. (6 points) Assume that $\sigma_k^2 = \sigma^2, \forall k$. An observation is classified to the class for which $P(Y = k|X = x)$ is the highest. Show that this is equivalent to

   $$\delta_k(x) = x\frac{\mu_k}{\sigma^2} - \frac{\mu_k^2}{2\sigma^2} + ln(\pi_k)$$

being the highest. $\delta_k(x)$ is called the discriminant function. $\pi_k = P(Y = k)$ (Hint: Use the Bayes Rule to find the expression of $P(Y = k|X = x)$.)

2. (8 points) Assume the more general case where $\sigma_1^2 \neq \sigma_0^2$. Calculate the Bayes' discriminant points analytically, ie: the points for which $P(Y = 1|X = x) = P(Y = 0|X = x)$. (Hint: Take the logarithm.)

3. (4 points) Assume that $\sigma_k^2 = \sigma^2, \forall k$. What is the ratio $\frac{P(X=x|Y=1)}{P(X=x|Y=0)}$ in the case of Gaussian densities?

4. (4 points) Assume equal prior, ie: $P(Y = 0) = P(Y = 1)$. How does the odds ratio obtained using the result of question (3) compared to the odds ratio of the logistic model?

5. (3 points) Comment the results obtained in question (4)

3 Exercise 3 (35 points): Who will open a bank account?

Demirgüç-Kunt et al. (2015) showed that only 54% of adults in developing countries report having a bank account. By contrast, only 6% of adults in the Organisation for Economic Co-Operation and Development (OECD) countries report not having a bank account.

This failure to access the banking system can act as a brake on economic and social development, as shown in the macroeconomic literature (Jayaratne and Strahan 1996; Black and Strahan 2002; Burgess and Pande 2005; Levine 2005; Beck, Demirgüç-Kunt, and Levine 2007; Bruhn and Love 2014).

To develop appropriate policies, the government must identify the population on which to act. Therefore, the Canadian government wants to deploy its expertise to help the Chilean government foster access to the banking sector in Chile. As an economist analyst at the Ministry of Foreign affairs of Canada, you are in charge of this task. Using a Chilean database, this exercise will guide you on how to do it.

1. (2 points) Read the file "banking.dta" in Python.

2. (2 points) Present descriptive statistics on the outcome variable B20 (having a bank account). Comment.

3. (10 points) Read carefully the description of the variables.

   - Choose the variables you will use to predict the likelihood of having a bank account.

- For each variable, explain why you think it matters and present some descriptives statistics (mean, median, standard deviation). (Hint: The goal is to make you thoughtfully choose variables, putting you in a real-life context. Then, you can use the knowledge of the field, the basic correlations, ... to build your choice. You will also notice that there are a lot of missing values (real-life data!). This should open up a discussion for future classes on how to handle them. But, at this point, I am just going to ask you to select your variables to have at the very least 1000 observations.)

4. (10 points) Construct the following classifiers using B20 as the outcome variable and the predictors are the variables you selected in question (3)

   - Logistic Classifier
   - KNN classifier (find optimal K and build the classifier)
   - Decision Tree Classifier
   - Random Forest classifier

5. (2 points) Comparing the area under the curve (AUC) criteria, find the best classifier among those in question (4).

6. (2 points) What are the top 3 most important features using the best classifier of question (5)?

7. (5 points) Using the decision tree, what are the top 3 characteristics of married individuals ("married individuals" is variable "n_C1_2") without a bank account? (Hint: check for tree interpreter)

8. (2 points) Based on what you have found, what is the key recommendation that you can make to the Chilean government that would like to foster the use of bank account among the population? (no more than 3 lines).

4 Exercise 4 (15 points): Webscrapping

Write an algorithm that does the following:

- Make a google query looking for the word : ”Machine Learning”
- Click and Open the first google search result. (Ideally, the first search result that you want to open with your algorithm should not be an ad.)

References:

Beck, Thorsten, Asli Demirgüç-Kunt, and Ross Levine. 2007. “Finance, inequality and the poor.” Journal of Economic Growth 12 (1): 27–49.

Black, Sandra E., and Philip E. Strahan. 2002. “Entrepreneurship and Bank Credit Availability.” Journal of Finance 57 (6): 2807–33.

Bruhn, Miriam, and Inessa Love. 2014. “The Real Impact of Improved Access to Finance: Evidence from Mexico.” Journal of Finance 69 (3): 1347–76.

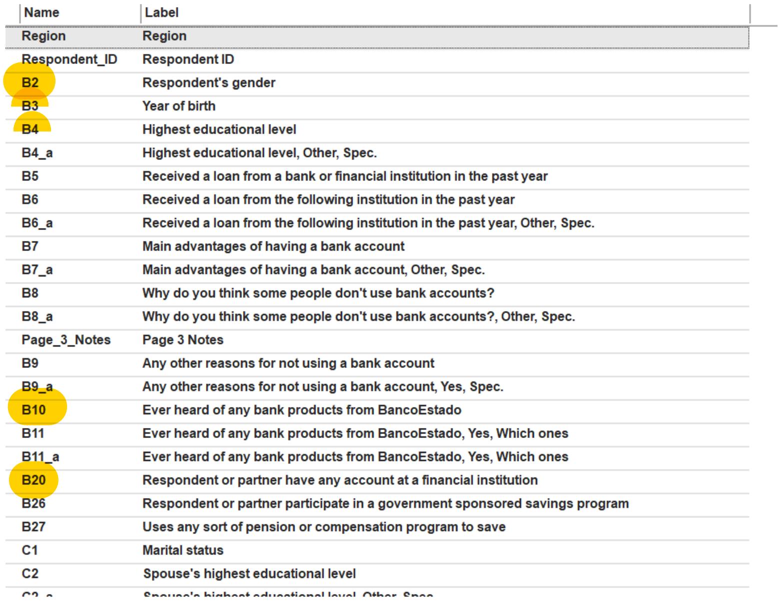
Burgess, Robin, and Rohini Pande. 2005. “Do Rural Banks Matter? Evidence from the Indian Social Banking Experiment.” American Economic Review 95 (3): 780–95.

Demirguç-Kunt, A., Klapper, L. F., Singer, D., & Van Oudheusden, P. (2015). The global findex database 2014: Measuring financial inclusion around the world. World Bank Policy Research Working Paper, (7255).

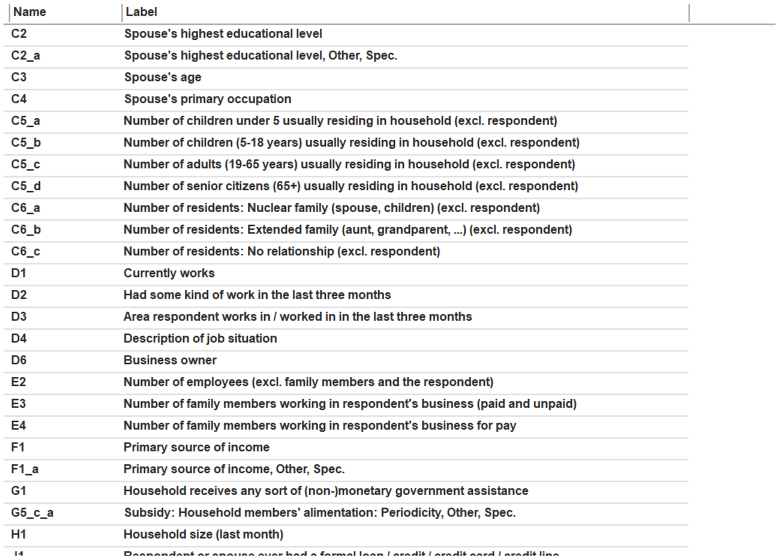
Dupas, Pascaline, Dean Karlan, Jonathan Robinson, and Diego Ubfal. “Banking the Unbanked? Evidence from Three Countries: Dataset.” American Economic Journal: Applied Economics.

Jayaratne, Jith, and Philip E. Strahan. 1996. “The Finance-Growth Nexus: Evidence from Bank Branch Deregulation.” Quarterly Journal of Economics 111 (3): 639–70.

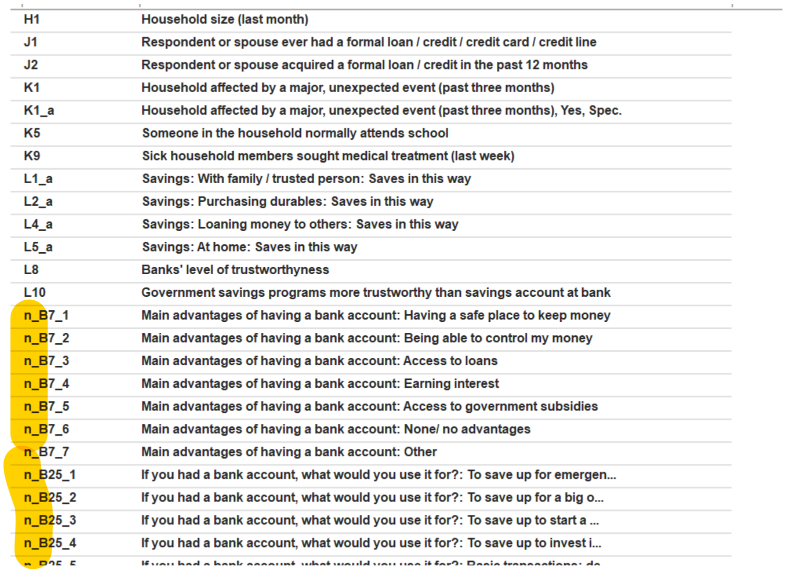
Variables Labels



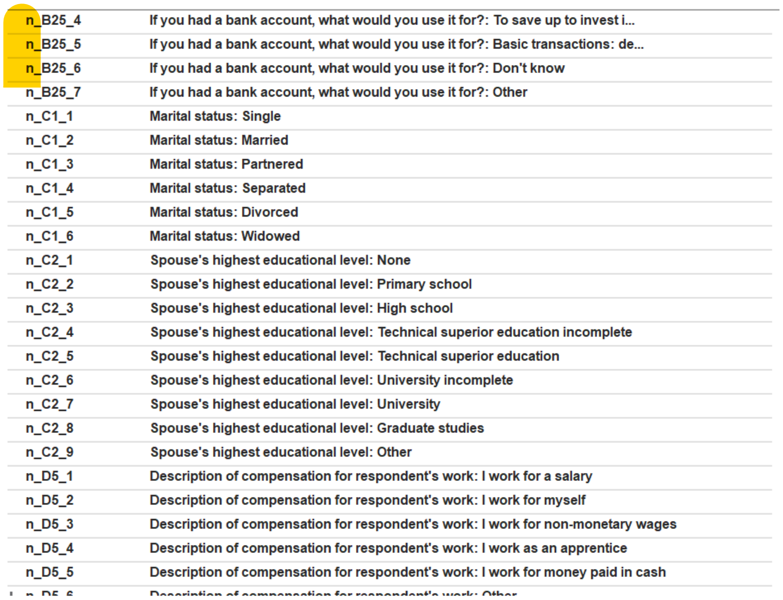
Name	Label
Region	Region
Respondent_ID	Respondent ID
B2	Respondent's gender
B3	Year of birth
B4	Highest educational level
B4_a	Highest educational level, Other, Spec.
B5	Received a loan from a bank or financial institution in the past year
B6	Received a loan from the following institution in the past year
B6_a	Received a loan from the following institution in the past year, Other, Spec.
B7	Main advantages of having a bank account
B7_a	Main advantages of having a bank account, Other, Spec.
B8	Why do you think some people don't use bank accounts?
B8_a	Why do you think some people don't use bank accounts?, Other, Spec.
Page_3_Notes	Page 3 Notes
B9	Any other reasons for not using a bank account
B9_a	Any other reasons for not using a bank account, Yes, Spec.
B10	Ever heard of any bank products from BancoEstado
B11	Ever heard of any bank products from BancoEstado, Yes, Which ones
B11_a	Ever heard of any bank products from BancoEstado, Yes, Which ones
B20	Respondent or partner have any account at a financial institution
B26	Respondent or partner participate in a government sponsored savings program
B27	Uses any sort of pension or compensation program to save
C1	Marital status
C2	Spouse's highest educational level
C2_a	Spouse's highest educational level, Other, Spec.



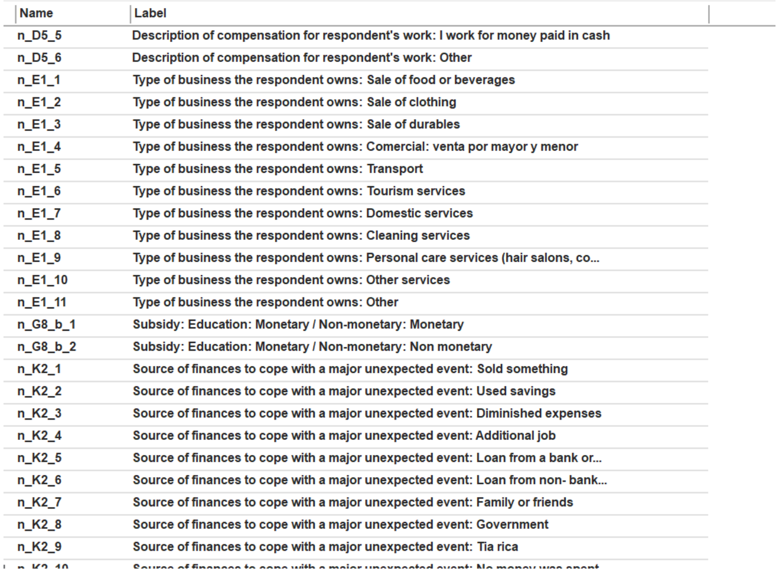
Name	Label
C2	Spouse's highest educational level
C2_a	Spouse's highest educational level, Other, Spec.
C3	Spouse's age
C4	Spouse's primary occupation
C5_a	Number of children under 5 usually residing in household (excl. respondent)
C5_b	Number of children (5-18 years) usually residing in household (excl. respondent)
C5_c	Number of adults (19-65 years) usually residing in household (excl. respondent)
C5_d	Number of senior citizens (65+) usually residing in household (excl. respondent)
C6_a	Number of residents: Nuclear family (spouse, children) (excl. respondent)
C6_b	Number of residents: Extended family (aunt, grandparent, ...) (excl. respondent)
C6_c	Number of residents: No relationship (excl. respondent)
D1	Currently works
D2	Had some kind of work in the last three months
D3	Area respondent works in / worked in in the last three months
D4	Description of job situation
D6	Business owner
E2	Number of employees (excl. family members and the respondent)
E3	Number of family members working in respondent's business (paid and unpaid)
E4	Number of family members working in respondent's business for pay
F1	Primary source of income
F1_a	Primary source of income, Other, Spec.
G1	Household receives any sort of (non-)monetary government assistance
G5_c_a	Subsidy: Household members' alimentation: Periodicity, Other, Spec.
H1	Household size (last month)
J1	Respondent or spouse ever had a formal loan / credit / credit card / credit line



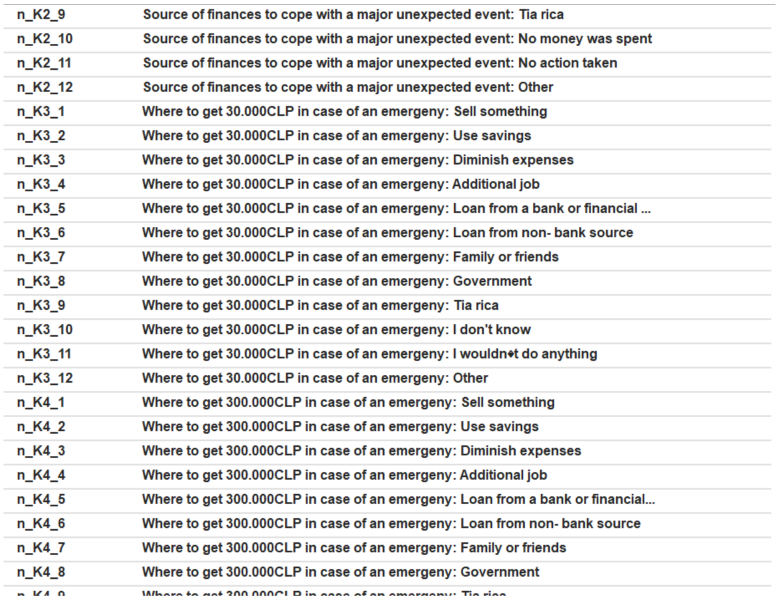
H1	Household size (last month)
J1	Respondent or spouse ever had a formal loan / credit / credit card / credit line
J2	Respondent or spouse acquired a formal loan / credit in the past 12 months
K1	Household affected by a major, unexpected event (past three months)
K1_a	Household affected by a major, unexpected event (past three months), Yes, Spec.
K5	Someone in the household normally attends school
K9	Sick household members sought medical treatment (last week)
L1_a	Savings: With family / trusted person: Saves in this way
L2_a	Savings: Purchasing durables: Saves in this way
L4_a	Savings: Loaning money to others: Saves in this way
L5_a	Savings: At home: Saves in this way
L8	Banks' level of trustworthyness
L10	Government savings programs more trustworthy than savings account at bank
n_B7_1	Main advantages of having a bank account: Having a safe place to keep money
n_B7_2	Main advantages of having a bank account: Being able to control my money
n_B7_3	Main advantages of having a bank account: Access to loans
n_B7_4	Main advantages of having a bank account: Earning interest
n_B7_5	Main advantages of having a bank account: Access to government subsidies
n_B7_6	Main advantages of having a bank account: None/ no advantages
n_B7_7	Main advantages of having a bank account: Other
n_B25_1	If you had a bank account, what would you use it for?: To save up for emergen...
n_B25_2	If you had a bank account, what would you use it for?: To save up for a big o...
n_B25_3	If you had a bank account, what would you use it for?: To save up to start a ...
n_B25_4	If you had a bank account, what would you use it for?: To save up to invest i...
n_B25_5	If you had a bank account, what would you use it for?: Basic transactions: de



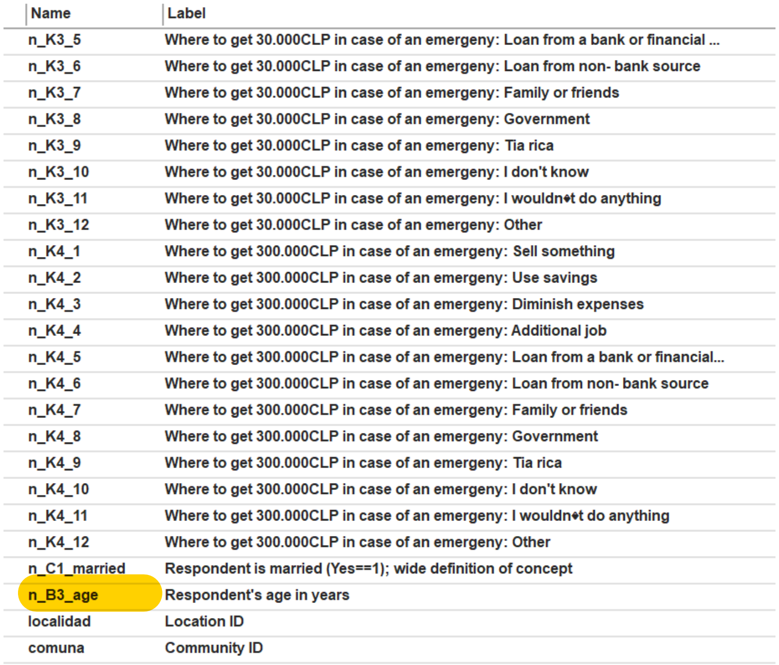
n_B25_4	If you had a bank account, what would you use it for?: To save up to invest i...
n_B25_5	If you had a bank account, what would you use it for?: Basic transactions: de...
n_B25_6	If you had a bank account, what would you use it for?: Don't know
n_B25_7	If you had a bank account, what would you use it for?: Other
n_C1_1	Marital status: Single
n_C1_2	Marital status: Married
n_C1_3	Marital status: Partnered
n_C1_4	Marital status: Separated
n_C1_5	Marital status: Divorced
n_C1_6	Marital status: Widowed
n_C2_1	Spouse's highest educational level: None
n_C2_2	Spouse's highest educational level: Primary school
n_C2_3	Spouse's highest educational level: High school
n_C2_4	Spouse's highest educational level: Technical superior education incomplete
n_C2_5	Spouse's highest educational level: Technical superior education
n_C2_6	Spouse's highest educational level: University incomplete
n_C2_7	Spouse's highest educational level: University
n_C2_8	Spouse's highest educational level: Graduate studies
n_C2_9	Spouse's highest educational level: Other
n_D5_1	Description of compensation for respondent's work: I work for a salary
n_D5_2	Description of compensation for respondent's work: I work for myself
n_D5_3	Description of compensation for respondent's work: I work for non-monetary wages
n_D5_4	Description of compensation for respondent's work: I work as an apprentice
n_D5_5	Description of compensation for respondent's work: I work for money paid in cash
n_D5_6	Description of compensation for respondent's work: Other



Name	Label
n_D5_5	Description of compensation for respondent's work: I work for money paid in cash
n_D5_6	Description of compensation for respondent's work: Other
n_E1_1	Type of business the respondent owns: Sale of food or beverages
n_E1_2	Type of business the respondent owns: Sale of clothing
n_E1_3	Type of business the respondent owns: Sale of durables
n_E1_4	Type of business the respondent owns: Comercial: venta por mayor y menor
n_E1_5	Type of business the respondent owns: Transport
n_E1_6	Type of business the respondent owns: Tourism services
n_E1_7	Type of business the respondent owns: Domestic services
n_E1_8	Type of business the respondent owns: Cleaning services
n_E1_9	Type of business the respondent owns: Personal care services (hair salons, co...
n_E1_10	Type of business the respondent owns: Other services
n_E1_11	Type of business the respondent owns: Other
n_G8_b_1	Subsidy: Education: Monetary / Non-monetary: Monetary
n_G8_b_2	Subsidy: Education: Monetary / Non-monetary: Non monetary
n_K2_1	Source of finances to cope with a major unexpected event: Sold something
n_K2_2	Source of finances to cope with a major unexpected event: Used savings
n_K2_3	Source of finances to cope with a major unexpected event: Diminished expenses
n_K2_4	Source of finances to cope with a major unexpected event: Additional job
n_K2_5	Source of finances to cope with a major unexpected event: Loan from a bank or...
n_K2_6	Source of finances to cope with a major unexpected event: Loan from non- bank...
n_K2_7	Source of finances to cope with a major unexpected event: Family or friends
n_K2_8	Source of finances to cope with a major unexpected event: Government
n_K2_9	Source of finances to cope with a major unexpected event: Tia rica
n_K2_10	Source of finances to cope with a major unexpected event: No money was spent



n_K2_9	Source of finances to cope with a major unexpected event: Tia rica
n_K2_10	Source of finances to cope with a major unexpected event: No money was spent
n_K2_11	Source of finances to cope with a major unexpected event: No action taken
n_K2_12	Source of finances to cope with a major unexpected event: Other
n_K3_1	Where to get 30.000CLP in case of an emergeny: Sell something
n_K3_2	Where to get 30.000CLP in case of an emergeny: Use savings
n_K3_3	Where to get 30.000CLP in case of an emergeny: Diminish expenses
n_K3_4	Where to get 30.000CLP in case of an emergeny: Additional job
n_K3_5	Where to get 30.000CLP in case of an emergeny: Loan from a bank or financial ...
n_K3_6	Where to get 30.000CLP in case of an emergeny: Loan from non- bank source
n_K3_7	Where to get 30.000CLP in case of an emergeny: Family or friends
n_K3_8	Where to get 30.000CLP in case of an emergeny: Government
n_K3_9	Where to get 30.000CLP in case of an emergeny: Tia rica
n_K3_10	Where to get 30.000CLP in case of an emergeny: I don't know
n_K3_11	Where to get 30.000CLP in case of an emergeny: I wouldn♦t do anything
n_K3_12	Where to get 30.000CLP in case of an emergeny: Other
n_K4_1	Where to get 300.000CLP in case of an emergeny: Sell something
n_K4_2	Where to get 300.000CLP in case of an emergeny: Use savings
n_K4_3	Where to get 300.000CLP in case of an emergeny: Diminish expenses
n_K4_4	Where to get 300.000CLP in case of an emergeny: Additional job
n_K4_5	Where to get 300.000CLP in case of an emergeny: Loan from a bank or financial...
n_K4_6	Where to get 300.000CLP in case of an emergeny: Loan from non- bank source
n_K4_7	Where to get 300.000CLP in case of an emergeny: Family or friends
n_K4_8	Where to get 300.000CLP in case of an emergeny: Government
n_K4_9	Where to get 300.000CLP in case of an emergeny: Tia rica



Name	Label
n_K3_5	Where to get 30.000CLP in case of an emergeny: Loan from a bank or financial ...
n_K3_6	Where to get 30.000CLP in case of an emergeny: Loan from non- bank source
n_K3_7	Where to get 30.000CLP in case of an emergeny: Family or friends
n_K3_8	Where to get 30.000CLP in case of an emergeny: Government
n_K3_9	Where to get 30.000CLP in case of an emergeny: Tia rica
n_K3_10	Where to get 30.000CLP in case of an emergeny: I don't know
n_K3_11	Where to get 30.000CLP in case of an emergeny: I wouldn�t do anything
n_K3_12	Where to get 30.000CLP in case of an emergeny: Other
n_K4_1	Where to get 300.000CLP in case of an emergeny: Sell something
n_K4_2	Where to get 300.000CLP in case of an emergeny: Use savings
n_K4_3	Where to get 300.000CLP in case of an emergeny: Diminish expenses
n_K4_4	Where to get 300.000CLP in case of an emergeny: Additional job
n_K4_5	Where to get 300.000CLP in case of an emergeny: Loan from a bank or financial...
n_K4_6	Where to get 300.000CLP in case of an emergeny: Loan from non- bank source
n_K4_7	Where to get 300.000CLP in case of an emergeny: Family or friends
n_K4_8	Where to get 300.000CLP in case of an emergeny: Government
n_K4_9	Where to get 300.000CLP in case of an emergeny: Tia rica
n_K4_10	Where to get 300.000CLP in case of an emergeny: I don't know
n_K4_11	Where to get 300.000CLP in case of an emergeny: I wouldn�t do anything
n_K4_12	Where to get 300.000CLP in case of an emergeny: Other
n_C1_married	Respondent is married (Yes==1); wide definition of concept
n_B3_age	Respondent's age in years
localidad	Location ID
comuna	Community ID