On the determinants of happiness: a Classification And Regression Tree (CART) approach.[☆]

Sergio Galletta[a,*]

[a] *Institute of Economics (IdEP), University of Lugano, 6904 Lugano, Switzerland*

Abstract

This article studies the determinants of the individual's subjective well-being by applying a classification and regression tree (CART) analysis to data from the Survey on Household Income and Wealth (SHIW) provided by the Bank of Italy. The results support the primary importance of economic conditions but show that their effect is heterogeneously dependent on other individual characteristics
< **Tables and figures at end** >
JEL Classification I31, C45, C49
Keywords Happiness, Classification trees, Italy

1. Introduction

In recent years, there has been a growing effort to understand the link between individual's happiness and economic conditions. For example, the declared level of happiness has been found to be correlated with several individual characteristics such as income (Clark et al., 2008) and employment status (Clark and Oswald, 1994), as well as with macroeconomic indicators like country inequality (Alesina et al., 2004; Ferrer-i Carbonell and Ramos, 2014), social capital (Rodríguez-Pose and Berlepsch, 2014), inflation and unemployment rate (MacCulloch et al., 2001). However, other personal aspects are also found to be important for happiness. For example, Vanassche et al. (2013) have found that the presence of children and the marital status affect individual satisfaction conditional to the country of residence.[1]

Typically these studies are based on estimations of standard parametric approaches such as OLS and ordered linear probit or logit regressions. These are common techniques that allow to identify the presence of correlation between a specified set of individual's features and his subjective well-being.

Regrettably, these models have a limited power to uncover multiple structures in those data which would suggest heterogeneity in the reported happiness within certain groups of individuals. Moreover they are not appropriate to deal with, otherwise preimposed, nonlinear relations between regressors and the dependent variable. Therefore, the purpose of this article is to provide results obtained using an alternative statistical methodology to complement previous research. To do so, I use a classification and regression tree (CART) approach to study the determinant of happiness of a sample of Italian citizens. CART is a data mining technique that, by identifying patterns in data, selects among a vector of explanatory variables those producing the best prediction of individuals' types (e.g., happy versus nonhappy).

More formally, trees are built through an algorithm that recursively partitions the data into nodes by iterated binary splits.[2] The root node (i.e., the whole sample) is therefore divided into other nodes (i.e.,

[☆] I wish to thank Massimo Bordignon and Ada Ferrer-i-Carbonell for insightful comments. Financial support from the Swiss National Science Foundation (grant Early Postdoc.Mobility - 158603) is gratefully acknowledged.
[*] Corresponding author.
Email address: sergio.galletta@usi.ch (Sergio Galletta)
[1] For a recent survey see Stutzer and Frey (2012).
[2] For a detailed technical explanation see Breiman et al. (1984).

subsamples) by following a set of rules which finds among all predictors the ones that allow for the most discriminative split. This is accomplished by testing the level of impurity of all possible splits. This procedure continues by creating branches and other nodes until certain conditions are met. Finally, the terminal nodes, or leaves, define the predicted type for each individual whose characteristics match the traced path. After a full decision tree is built, there is usually the need to prune some branches. Indeed, by pruning the tree the results are easier to understand as well as more precise in classifying alternative data-sets (Han et al., 2011).

Although this statistical method has been used occasionally in economics (see, for example, Manasse and Roubini (2009), Keely and Tan (2008) and Williams et al. (1987)), classification algorithms are used frequently in several disciplines such as medicine, engineering, marketing and biology. To the best of my knowledge, this is the first work showing the use of a classification tree to study the potential determinants of happiness.

2. Data

To create the classification tree, I use data from the last five waves of the Survey on Household Income and Wealth (SHIW) conducted by the Bank of Italy every 2 years. Hence, I take into account data from the years 2004, 2006, 2008, 2010 and 2012.[3] By merging the data-sets from these waves I ended up with a sample consisting of 14,158 individuals. Of them 3,026 are present in more than one wave. Therefore, I avail of 20,243 independent individual-year observations.

The questions asked in the survey are concerned mainly with household's economic and sociodemographic backgrounds. However, and interestingly for the aim of the article, the survey has a section where it asks householders how happy they are with their life through the following question: *"Considering all the aspects of your life, how happy would you say you are? Please score on a scale from 1 to 10, where 1 means extremely unhappy and 10 extremely happy, and the intermediate numbers serve to graduate"*.Answers from this kind of question are typically considered a good proxy of the level of subjective well-being.[4]

Given this information, I derive the dependent variable *happy*, which is a dummy variable equal to 1 if an individual reports a level of happiness higher than the sample average (6.9), and 0 otherwise. Figure 1 shows the sampling distribution of this variable. Roughly 65% of the sample is happier than the average while 35% is less happy.

I use several explanatory variables that identify different aspects of the individual's life. Table 1 lists and describes in more detail these variables, which for the sake of clarity I divide into five groups: *socio-demographic*, *work*, *income*, *consumption* and *wealth* status.[5]

It is worth noting that, compared with previous literature, I use a larger number of covariates, 25 to be exact. This is possible because of the flexibility that classification trees feature in handling large set of potentially interrelated variables. To put this differently, to run a standard regression by including the same amount of information, because of the use of categorical variables, one would have to include 46 regressors. Moreover, I include together both aggregate measure variables and their main components. For example, I will test whether general *consumption* is important to determine the individual's happiness but at the same time I try to discriminate which type of consumption is important (i.e., durable versus nondurable goods). Same argument holds for *income* and *wealth*.

[3] Scoppa and Ponzo (2008) use SHIW from the years 2004 and 2006 producing the first empirical study on happiness focused on Italy.

[4] Alesina et al. (2004) in section 2.2 discuss in more details this point.

[5] Few monetary variables report negative values. This is due to the way in which the Bank of Italy aggregates the different components. However, it is worth to note that even if these values were misreported data, it would not be an issue because classification trees would handle it efficiently (Breiman et al., 1984).

3. Empirical procedure and findings

The following text describes the main settings I use to build the classification tree.[6] First, I randomly select four-fifth of individuals (i.e., "training sample") to grow the tree, while I use the remaining one-fifth of the sample (i.e., "testing sample") to test the precision of the model predictions. Second, after constructing the full tree without imposing any restriction on the minimum size of the node, I prune it by applying a *cost complexity* algorithm. This process minimizes the misclassification error by taking into account a penalty for additional nodes. I use the *Gini index* to measure the level of node impurity. Formally, $Gini_{index} = 1 - \sum_{i=1}^{m} p_i^2$, where p_i is the probability that an individual i from the sample belongs to one of the classes identified in the dependent variable. This is the standard criterion used in classification trees.

Figure 2 shows the best classification tree by including the discriminant variables, the probability that an individual in each node reports a level of happiness higher than the national average and the predicted type just for the terminal nodes.[7] The pruned tree has a mis-classification error of nearly 30%.

Among the 25 explanatory variables the tree selected just 8 of them. These variables are: *net disposable income*, *payroll income*, *financial assets*, *real assets*, *area of residence*, *age class*, *branch of activity* and *education*. Interestingly, there is at least one variable for each of the five groups previously identified, confirming that subjective well-being involves many aspects of life.

The first split depends on the household's level of *income*: households with an high income (i.e., higher than 27,000 euro) go to the right branch, and the others go to the left. This first split suggests that people with high income have 24% higher probability to be classified as *happy* than people with a low income. The next splitting variable, in both branches, is the *marital status* .

On the right branch, those that are married have 81% probability of being *happy*, while all nonmarried (i.e., single, separated/divorced and widow/er) have a 61% probability to declare themselves *happy*. However, single and divorced individuals are eventually separated from widowers, who report a lower probability to be *happy*. For the latter, the value of *real asset* plays a significant and positive role to determine their happiness.

On the left branch, single and married people are grouped against divorced and widowed. The first group is further divided depending on *income*. For those with an income higher than 16,000 euro the likelihood of being happy is 17% higher than those with an income lower than this threshold. However, also the *sector of activity* and the *age* are involved in the classification of this subsample. Unemployed and people working in the industry are less likely to be happy compared with people working in other sectors. Also younger people do better. The second group instead (i.e., divorced and widower) is the one reporting aamong the lowest probability of being happy: the likelihood of being happy here is 25% lower than the whole sample. Further, the circumstances that affect happiness for this subsample are the living *area*, *payroll income* and *wealth*. Indeed, living in the centre or in the north of Italy, declaring an high payroll income and a high wealth from financial assets favour happiness. Strikingly, and in spite of what has already been found, a group of individuals from this branch has a decreased probability of being happy though they have a high wealth from real assets.

4. Discussion

The empirical results suggest that happiness depends positively on income as every time it appears in the tree the sample is split such that the likelihood of an individual to be happier than the national average is greater for higher values of this variable. Financial assets have the same features of income while, counter-intuitively, real assets both reduce and increase the likelihood of being happy depending on the group considered. However, other noneconomical features are important and among those the marital status plays a central role. Indeed, being married positively affects happiness regardless of the level of income. Although the tree does not allow to produce any clear hypothesis testing because there are no results in terms of inference, these findings give a descriptive insight coherent with previous research.

[6] All results provided in this paper are derived by using the *rpart* module implemented in **R**.

[7] The predicted type is *happy* if the probability is higher than 50%.

Interestingly, however, by using a classification tree instead of a standard parametric approach, I emphasize the heterogeneous pattern of the likelihood of being happy. For example, there are individuals from the main left branch (i.e., income $<$ 27,000 euro) that have a higher probability to be happy compared to individuals from the right branch (i.e., income $>$ 27,000 euro). Indeed, by looking at the ranges of probability of the 13 terminal nodes, grouped according to the first split (4 and 9, respectively, for income higher and lower than 27,000), the relevance of interactions between variables becomes more clear. More than 75% of nodes from the poorer group have a likelihood of being happy equal or higher than the group of richest people with the lowest probability of being happy. Same reasoning is possible by considering other branches.

In conclusion, CART seems to perform well. It reveals how variables interact with each other without imposing any model specification. This helps to show that heterogeneity in responses to changes of status is relevant for studies on happiness. Therefore, researchers should carefully take it into account every time they are conscious it might affect theirs results.

References

Alesina, A., Tella, R. D., MacCulloch, R., 2004. Inequality and happiness: are europeans and americans different? Journal of Public Economics 88 (9–10), 2009 – 2042.

Breiman, L., Friedman, J. H., Olshen, R. A., Stone, C. J., 1984. Classification and Regression Trees. Chapman & Hall, New York, NY.

Clark, A. E., Frijters, P., Shields, M. A., March 2008. Relative Income, Happiness, and Utility: An Explanation for the Easterlin Paradox and Other Puzzles. Journal of Economic Literature 46 (1), 95–144.

Clark, A. E., Oswald, A. J., May 1994. Unhappiness and Unemployment. Economic Journal 104 (424), 648–59.

Ferrer-i Carbonell, A., Ramos, X., 2014. Inequality and happiness. Journal of Economic Surveys 28 (5), 1016–1027.

Han, J., Kamber, M., Pei, J., 2011. Data Mining: Concepts and Techniques, 3rd Edition. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

Keely, L. C., Tan, C. M., June 2008. Understanding preferences for income redistribution. Journal of Public Economics 92 (5-6), 944–961.

MacCulloch, R. J., Tella, R. D., Oswald, A. J., March 2001. Preferences over Inflation and Unemployment: Evidence from Surveys of Happiness. American Economic Review 91 (1), 335–341.

Manasse, P., Roubini, N., July 2009. "Rules of thumb" for sovereign debt crises. Journal of International Economics 78 (2), 192–205.

Rodríguez-Pose, A., Berlepsch, V., April 2014. Social Capital and Individual Happiness in Europe. Journal of Happiness Studies 15 (2), 357–386.

Scoppa, V., Ponzo, M., June 2008. An Empirical Study of Happiness in Italy. The B.E. Journal of Economic Analysis & Policy 8 (1), 1–23.

Stutzer, A., Frey, B. S., Dec 2012. Recent Developments in the Economics of Happiness: A Selective Overview. IZA Discussion Papers 7078, Institute for the Study of Labor (IZA).

Vanassche, S., Swicegood, G., Matthijs, K., April 2013. Marriage and Children as a Key to Happiness? Cross-National Differences in the Effects of Marital Status and Children on Well-Being. Journal of Happiness Studies 14 (2), 501–524.

Williams, M. A., Joskow, A. S., Johnson, R. L., Hurdle, G. J., 1987. Explaining and predicting airline yields with non-parametric regression trees. Economics Letters 24 (1), 99–105.

Table 1: Variables description

Variable name	Values
Dependent variable	
Happy	avg=0.64, min=0, max=1
Socio-demographic	
Sex	male, female
Marital status	married, single, separated/divorced and widow/er
Education	none, primary school certificate,lower secondary school certificate, vocational secondary school diploma (3 years of study), upper secondary school diploma, 3-year university degree/higher education diploma, 5-year university degree, postgraduate qualification
Area of birth	north, centre, south and islands, abroad
Area of residence	north, centre, south and islands
Age class	Up to 30 years, 31-40, 41-50, 51-65, more than 65 years
Household members	avg=2.48, min=1, max=11
Living with children	yes, no
Working life	
Employment status	blue-collar worker, office worker or school teacher, cadre or manager, sole proprietor/member of the arts or professions,other self-employed, pensioner, other not-employed
Sector of activity	agriculture, industry, public administration, other sector, not employed
Jobs in life	avg=1.90, min=0, max=40
Income	
Net disposable income	avg=34573, min=-49558, max=1218959
Payroll income	avg=12524, min=0, max=247250
Pensions and net transfers	avg=10218, min=-28608, max=494235
Net self-employment income	avg=4300, min=-59600, max=1192000
Property income	avg=7528, min=-21358, max=233038
Consumption	
Consumption	avg=26341, min=-7077, max=307914
Durables Cons.	avg=1562, min=-33674, max=273750
Non-durables Cons.	avg=24779, min=1820, max=264960
Valuables Cons.	avg=85, min=0, max=35760
Means of transport Cons.	avg=883, min=4875, max=873750
Household appliances Cons.	avg=534, min=2013, max=46000
Wealth	
Net wealth	avg=271958, min=-775494, max=31274192
Real assets	avg=253057, min=0, max=31274192
Financial assets	avg=29746, min=0, max=4883130
Financial liabilities	avg=9998, min=0, max=4301000

All monetary values are in euro at 2013.

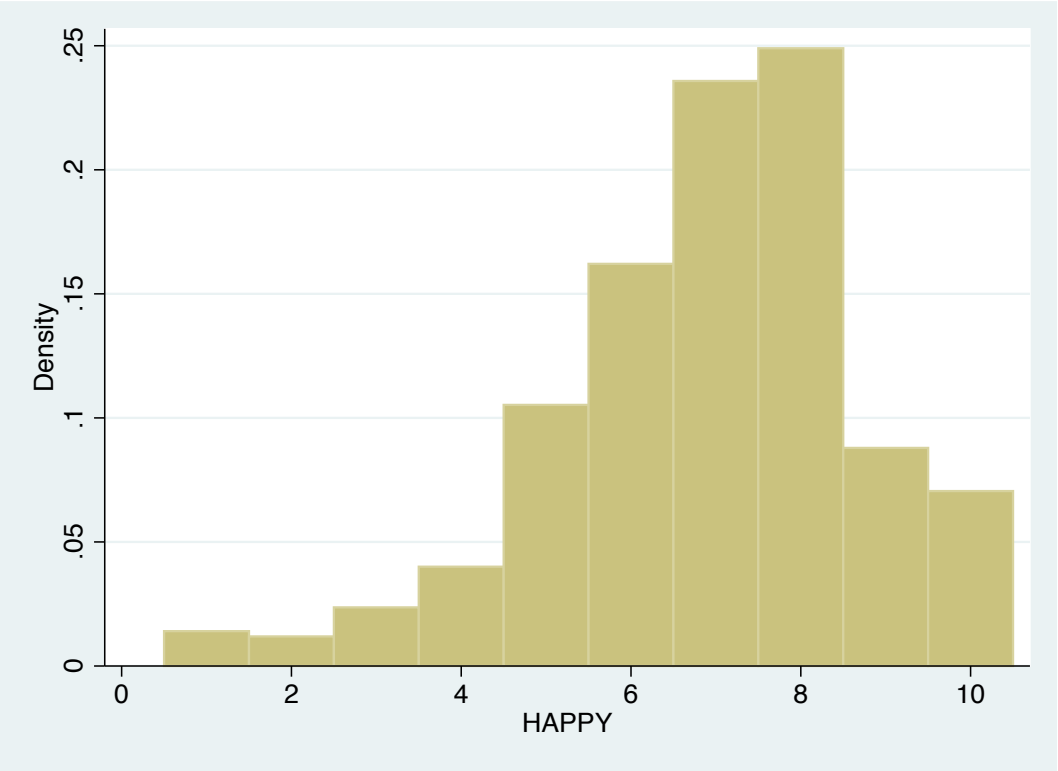

Figure 1: Distribution happiness.

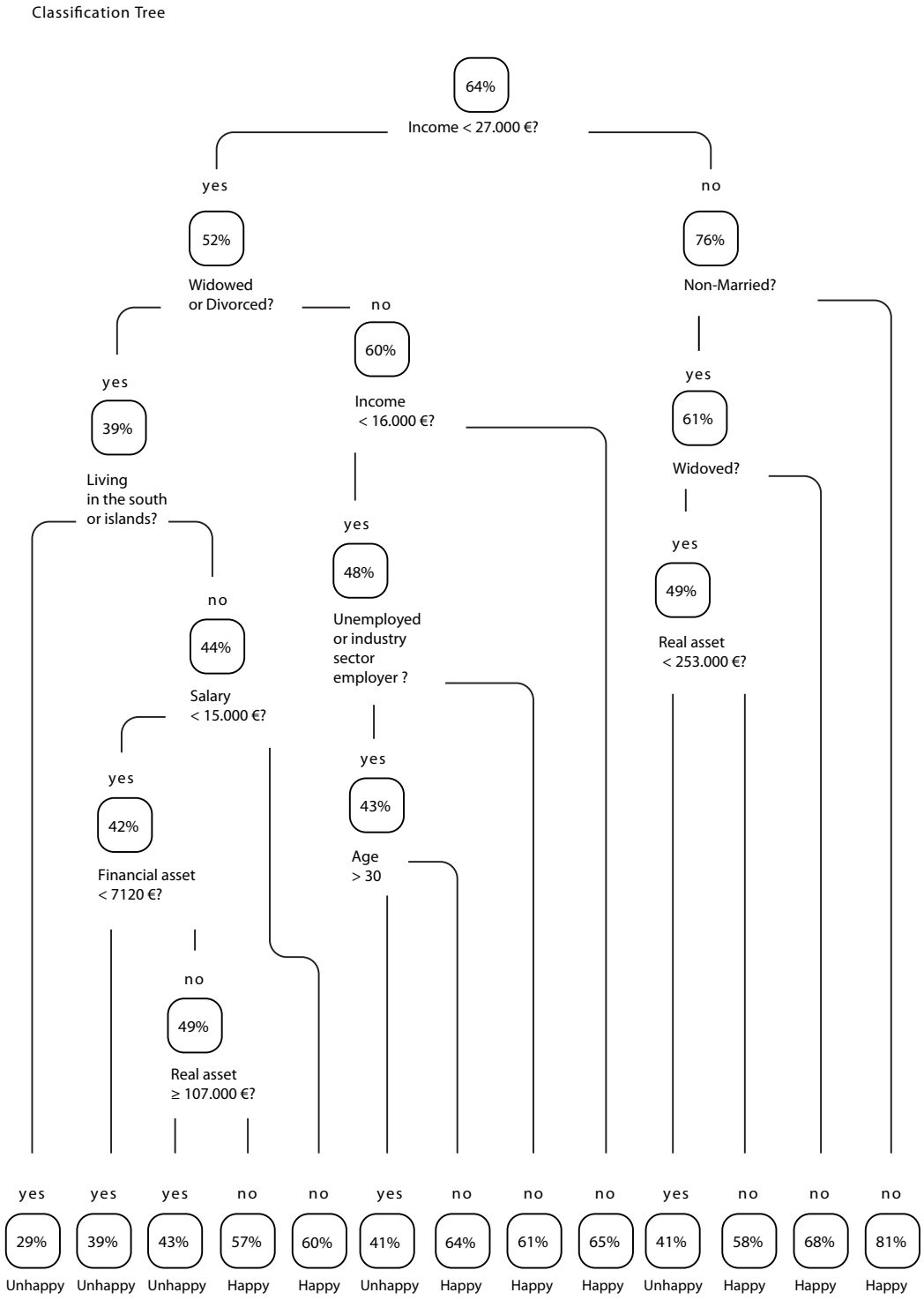

Figure 2: The Classification tree.