Universiteit Utrecht

Master thesis Applied Data Science at Utrecht University

Wiktoria Niewiadomska, 0993972

Topic: Predicting wellbeing in Latin America

Thesis supervisors:

Dr. Yolanda Grift

Dr. Tina Dulam

Abstract

Individual life satisfaction has implications that go beyond one's personal life having potential to influence social and economic indicators as well. Its significance has made it a field of study in recent decades. However, these studies are normally focused on the Western world. Very few look at other regions despite evidence pointing to cultural differences in what influences life satisfaction. This paper takes a closer look at subjective wellbeing in Latin America. This region has been found to be happier than what economic indicators alone would suggest, making it an interesting object of study. This piece employs supervised learning methods to determine whether currently known life satisfaction predictors are enough to reliably predict wellbeing. It also takes a look at relative importance of its predictors in this way adding research with the use of novel methods to the limited body of knowledge on the topic.

Introduction, motivation and context

In recent decades, life satisfaction and its measurements have become a topic of scientific inquiry. High wellbeing has been linked to success on personal, interpersonal and professional levels. It has been determined to lead to higher productivity, better learning, creativity, pro-social behaviors as well as positive relationships (Ruggeri et al., 2020). Therefore, individual's life satisfaction has implications that go way beyond one's personal life, making it a subject of study in fields ranging from psychology to economics and warranting the increased interest in the topic manifested by the emergence of journals and measures focused on happiness and life satisfaction alone. Just as in majority of fields, however, the focus of such studies is usually what came to be known as the "Western world". According to Tov and Au (2013), only about 30% of all empirical research in this field concerns countries outside Europe and North America (Tov & Au, 2013). Latin America specifically, despite its growing economic and social significance, has only been researched in a few studies. It was found to exhibit some interesting features not applicable to other regions. To be precise, while the level of individual economic wellbeing is seen as a major factor in happiness in majority of the world, in Latin America its influence can be moderated by participation in religious services, strength of social ties as well as interpersonal trust (Conci, 2017). Additionally, the example of Latin America goes against the well-established Easterlin paradox making the region an interesting study subject (Corral, 2011)

Literature overview

Academic literature identifies numerous predictors of subjective wellbeing. In an extensive review of research on the topic, Dolan et al. (2008) have found that subjective well-being has been linked to health and employment status and social contacts. One of the other important predictors is considered to be the perception of one's financial status (Oishi et al., 2009; Diego-Rosell et al., 2016). This factor is more important in poorer nations, while in these more well-off social factors may take precedence (Oishi et al., 2009). It has also been found that such determinants differ between individualist and collectivist cultures, which highlights that the research on the Western world is not universally applicable and indicates the need to conduct research in other regions. Other research has found that both job and personal life (family) satisfaction influence general happiness levels (Near et al., 1984), pointing to the importance of both professional and private life to happiness. At the same time, romantic relationships and religiosity have been determined in a recent study to be able to prevent one from happiness drops in times of crisis indicating a cushioning moderating effect (Stone, 2022).

Studies specifically looking at the Latin American countries have identified a number of additional predictors. Singer (2013) discovered the effect of prevalence of bribery on life satisfaction and Ortega Londoño et al. (2019) estimated the negative influences of crime and victimization. It has also been found that local governments services have an influence on the perception of satisfaction. Another important factor has been religion, that has been found to mitigate negative effects of other factors, such as unemployment (Garzon & Ruprah, 2015). Some research focused on the United States and Latin America also pointed to differences between rural and urban communities, with the former were found to be happier in the United States but no different in Latin America (Valente & Berry, 2016). There, social ties, like relationships with family and friends, have been determined to be more influential.

Research question

The literature review above points to a number of determinants of wellbeing from a number of domains spanning personal, social and professional life as well as influence of reliability of government services and support. It has also pointed to differences between cultures and the fact that in the limited number of previous studies Latin American countries sometimes present different wellbeing determinants that Western states. At the same time, despite lower economic indicators, Latin American states rank just after Northern America and Western Europe in terms of happiness, with some studies finding that Latinos are happier than modeling based on economic factors would suggest (Conci, 2017). Therefore, looking at subjective well-being in this region can bring new insights regarding life satisfaction across cultural boundaries. This paper aims to find whether a model with currently known factors impacting life satisfaction can accurately predicting it in Latin American countries. It does so by applying methods of supervised learning to results of the comprehensive 2016-17 wave of LAPOP AmericasBarometer survey, a representative picture of attitudes of inhabitants of the region. This study also aims to find the relative importance of factors relating to wellbeing in the region. In this way, it will enable comparisons between Latin America and other regions.

Data

The original 2016-17 LAPOP dataset consists of 42,451 observations across 535 variables spanning areas such as demographics, socioeconomic status, various aspects of life satisfaction (from general to satisfaction with infrastructure or government), social and political participation and views (Vanderbilt University, 2017). The dataset included 32 countries in Latin America and 2 North American ones. Based on findings of Dulam, Grift & Van den Berg (2021), observations from Venezuela and Haiti, outliers due to their exceptionally low wellbeing levels, were removed from analysis. The same approach was applied to North American countries, which are not a subject of this study.

For the purpose of this study, a subset of 31 variables from the original dataset was chosen. These variables were included based on theoretical relevance i.e. previous studies in the field, as summarized in literature review. Most variables, 27 out of 31, owing to the qualitative natures of questions, are categorical. Out of them, one is an ordinal measure of subjective wellbeing – the outcome variable. The leftover minority, 4 variables of mainly demographic type, is numeric. The continuous variables were converted to numeric type and the categorical were saved as factors. Observations with missing variables were removed resulting in the final dataset of size 18,200 rows. Additionally, outcome variable (response to: “In general, how satisfied are you with your life?”) was transformed from four levels to two by pooling “somewhat (dis)satisfied” and “very (dis)satisfied” groups together. This was done to avoid classifying on very small samples, since the negative groups were underrepresented, accounting for just 13% of the sample (Table 1). Analysis of correlation between the variables has not revealed anything unexpected. Variables were strongly correlated in thematic groups. For example, social development variables were connected with each other as were these relating to liberties and freedoms. The similar variables were included as they captured different aspects of theoretically relevant domains affecting life satisfaction. By feature selection methods, these contributing little to the predictions were later removed. Variance inflation factor, run after model fitting, did not reveal multicollinearity. The descriptive statistics of included observations are shown in Table 1 below.

Table 1 – Summary statistics

Variable	**N**	**Mean**	**Std. Dev.**	**Min**	**Pctl. 25**	**Pctl. 75**	**Max**
age	18200	37.881	15.514	16	25	48	112
years of education	18200	9.84	4.175	0	7	12	18
# of children	18200	2.044	2.097	0	0	3	20
# of children in the house	18200	1.165	1.27	0	0	2	13

country	18200	
Mexico	1275	7%
Guatemala	1184	6.50%
El Salvador	1356	7.50%
Honduras	1267	7%
Nicaragua	1281	7%
Costa Rica	1269	7%
Panama	1297	7.10%
Colombia	1311	7.20%
Ecuador	1270	7%
Peru	2269	12.50%
Paraguay	920	5.10%
Brazil	1346	7.40%
Dom. Rep.	1095	6%
Jamaica	1060	5.80%
financial status	18200	
… Good enough and can save	1751	9.60%

… Good enough, with no major problems	6663	36.60%
… Not enough, and are stretched	6252	34.40%
… Not enough, and having a hard time	3534	19.40%
religious services attendance	18200	
… More than once a week	3062	16.80%
… Once a week	4686	25.70%
… Once a month	3559	19.60%
… Once or twice a year	2829	15.50%
… Never or almost never	4064	22.30%
people in the community…	18200	
… Very Trustworthy	4525	24.90%
… Somewhat Trustworthy	5982	32.90%
… Not Very Trustworthy	5609	30.80%
… Untrustworthy	2084	11.50%
sex	18200	

… Male	9358	51.40%
… Female	8842	48.60%
occupation	18200	
… Working	7837	43.10%
… Not working at the moment, but have a job	1139	6.30%
… Actively looking for a job	2485	13.70%
… Student	1475	8.10%
… Taking care of the home	3600	19.80%
… Retired, pensioner or permanently disabled to work	933	5.10%
… Not working and not looking for a job	731	4%
size of municipality	18200	
… Large	7841	43.10%
… Medium	5264	28.90%
… Small	5095	28%
attended a town meeting	18200	

… Yes	2363	13%
… No	15837	87%
safety in neighborhood	18200	
… Very Safe	3782	20.80%
… Somewhat Safe	6490	35.70%
… Somewhat Unsafe	4766	26.20%
… Very Unsafe	3162	17.40%
victim of crime last year	18200	
… Yes	4554	25%
… No	13646	75%
condition of roads	18200	
… Very Satisfied	1251	6.90%
… Satisfied	7936	43.60%
… Dissatisfied	7027	38.60%
… Very Dissatisfied	1986	10.90%
quality of schools	18200	
… Very Satisfied	1732	9.50%

… Satisfied	9753	53.60%
… Dissatisfied	5232	28.70%
… Very Dissatisfied	1483	8.10%
quality of health services	18200	
… Very Satisfied	1264	6.90%
… Satisfied	6719	36.90%
… Dissatisfied	7277	40%
… Very Dissatisfied	2940	16.20%
attention to the news	18200	
… Daily	11266	61.90%
… A few times a week	4531	24.90%
… A few times a month	436	2.40%
… Rarely_never	1967	10.80%
freedom of press	18200	
… Very Little	9127	50.10%
… Enough	4780	26.30%
… Too Much	4293	23.60%

freedom of political expression	18200	
… Very Little	11099	61%
… Enough	4681	25.70%
… Too Much	2420	13.30%
human rights protection	18200	
… Very Little	12481	68.60%
… Enough	4039	22.20%
… Too Much	1680	9.20%
economic sit. vs. last year	18200	
… Better	3565	19.60%
… Same	7559	41.50%
… Worse	7076	38.90%
“Those who govern this country are interested in what people think”	18200	
… Strongly Disagree	4393	24.10%
… 2	2280	12.50%

… 3	2388	13.10%
… 4	2717	14.90%
… 5	2728	15%
… 6	1736	9.50%
… Strongly Agree	1958	10.80%
relationship status	18200	
… Single	6472	35.60%
… Married/civil union	5486	30.10%
… Common law marriage (living together)	4596	25.30%
… Divorced	389	2.10%
… Separated	681	3.70%
… Widowed	576	3.20%
protest participation	18200	
… Yes, I have participated	1737	9.50%
… No, I have not participated	16463	90.50%
Political knowl. of respondent	18200	

… high	6445	35.40%
… Neither High Nor Low	8217	45.10%
… low	3538	19.40%
interest in politics	18200	
… A lot	2163	11.90%
… Some	3781	20.80%
… Little	5988	32.90%
… None	6268	34.40%
voted in presidential elections	18200	
… Voted	13034	71.60%
… Did not vote	5166	28.40%
police officer asked you for a bribe	18200	
… No	15794	86.80%
… Yes	2406	13.20%
government employee ask you for a bribe	18200	

… No	16978	93.30%
… Yes	1222	6.70%
satisfaction	18200	
… 0 (dissatisfaction)	2357	13%
… 1 (satisfaction)	15843	87%

In the modeling stage, the dataset was split into training and testing. To avoid issues related to training on samples with imbalanced outcome variable distribution, the training set was a stratified sample consisting of 80% of 0-level variables (n=2004) with the same number sampled from 1-level observations resulting in a total training set of 4008 observations. The practice of limiting the number of observations from the more common class is known as down-sampling and is widely accepted in handling imbalanced data (Lee & Seo, 2022). The train set was used to fit the models, when test set was used for testing the predictions developed based on the fitted train model. These predictions were later used to compute accuracy metrics for each of the models.

Data preparation for analysis

The LAPOP dataset comes in two forms: country-level or integrated. For this analysis, the integrated dataset was used. Data exploration revealed a high number of missing observations. 18,200 were found to be complete cases with regards to variables in question.

The literature on the topic identifies three main types of missing values: missing completely at random (MCAR), missing at random (MAR) and missing not at random

(MNAR) (Bhaskaran & Smeeth, 2014). The case of MCAR is the best scenario for listwise deletion, yet, at the same time, it is a very uncommon one. Normally, missingness is dependent on other factors, whether they are (MAR) or are not (MNAR) present in the dataset (Van Buuren, 2018). To employ the appropriate method of missing data handling, the analysis of missing data was performed (Table 2).

Table 2 – Overview of most commonly missing variables

variable	**# missing**	**% missing**
attendance of town hall meetings (past year)	6421	22.59483
quality of schools	1180	4.152298
freedom of press	854	3.005138
quality of health services	695	2.445633
trust in community members	684	2.406925
human rights protection	599	2.107819
satisfaction with income	585	2.058554
freedom of political expression	578	2.033922
years of education	514	1.808713

“Those who govern this country are interested in what people think”	506	1.780562
safety in neighborhood	335	1.17883
quality of roads	293	1.031037

*only variables with >1% of missing observations included

The analysis revealed that missingness is a problem that affects one variable particularly strongly (np1, attendance of town hall meetings in past year). The number and share of missing values per variable is summarized in table above. The figures suggest that the problem does not affect variables in a significant manner, with the exception of the first listed variables. Numbers alone, however, tell us little about the origins of missingness as well as missingness pattern. Hence, two more figures were plotted to take a closer look at these factors.

Figure 1 – Missingness pattern

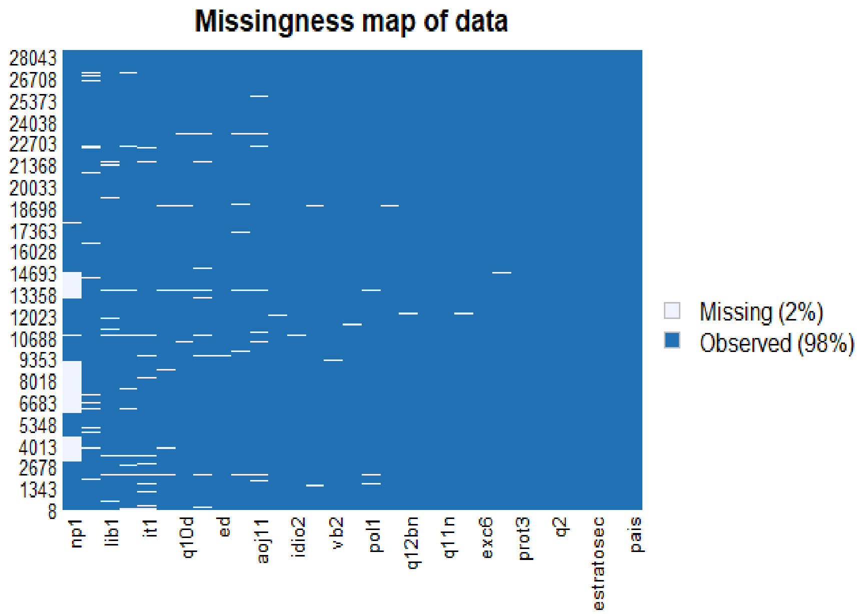
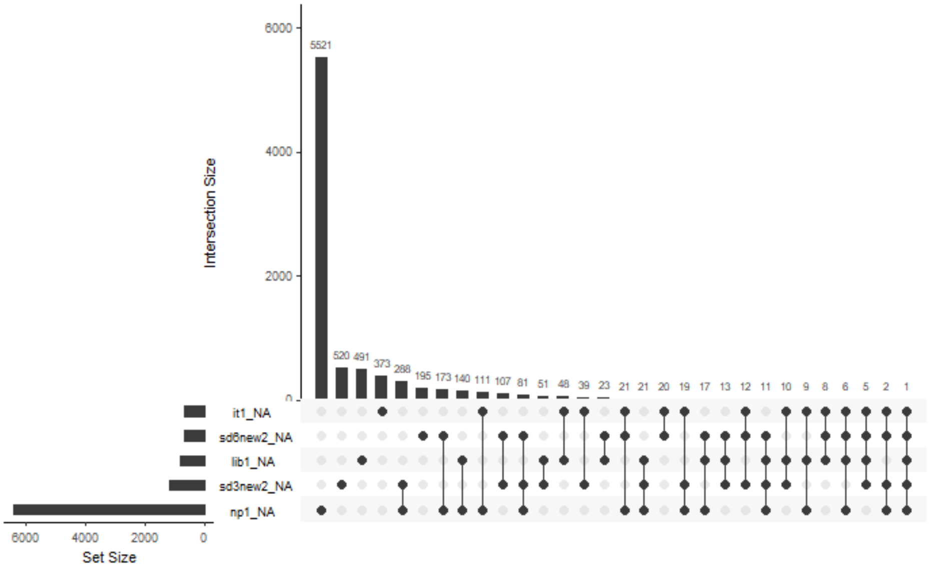


Figure 2 – Correlations between missing variables



While Figure 1 shows a few cases when multiple variables are missing together, Figure 2 reveals that this is not a common situation. There does not seem to be a significant number of cases, considering sample size, when variables miss together. Therefore, it does not appear that there is a pattern in which variables tend to miss. To confirm that, statistical analysis was performed. There is no statistically significant link between missingness across other variables and the most commonly missing variable. More importantly, in none of the variables missingness was significantly related to the outcome variable (included in Appendix). This points to one special case when use of listwise deletion, even under MAR or MNAR, can be appropriate (Van Buuren, 2018). Taking this into account, it appears that complete case analysis could be an appropriate method to proceed with the analysis, since removing them will not significantly change analysis results. Since only complete cases are used, cases of analysis performed on varying subsamples are avoided, hence removing inconsistency – another danger of listwise deletion. Therefore, this simple and robust method of dealing with missing data was chosen.

Ethical and legal considerations of the data

To the best of author's knowledge, there are no ethical or legal implications related to the use of this dataset. The used data was made available by LAPOP Lab and its major supporters: the United States Agency for International Development, the Inter-American Development Bank, and Vanderbilt University. The dataset used in this study is the 2016-2017 wave available at the website of the project (*Data Sets*, n.d.).

Methods

Translation of the research question to a data science question

This paper aims to contribute to research on predictors of subjective wellbeing in Latin America. It builds upon previously identified predictors of wellbeing and uses methods of supervised learning to predict a binary outcome of life satisfaction (1) or dissatisfaction (0). It does so with the use of logistic regression, a statistical technique commonly used in social sciences, and Support Vector Classification, which is used more rarely, but generally seen as a reliable and effective classification model (Zhang, 2017). It also makes use of two feature selection algorithms, the results of which motivate the inclusion of variables into tested models by providing values of information gain related to the variables. With these methods, the study is able to answer whether the currently known wellbeing predictors are sufficient to provide a reliable prediction of life satisfaction, as well as determine how important predictors are relative to each other.

Selection of methods for analysis

To evaluate the importance of variables, two methods of variable selection were used: FSelector and Boruta from R software. Considering the nature of predictors – mostly categorical – and outcome – categorical as well – the method used within FSelector was information gain (Romanski et al., 2021). There were two models chosen for evaluation. One included all variables bringing about information gain greater than 0. Another was the optimal subset chosen by the algorithm.

Boruta was used at its standard setting i.e. using Random Forest to perform a search for relevant features by comparing their estimated importance with what would be achieved at random (Kursa & Rudnicki, 2010). Based on this, Boruta classifies variables as important, unimportant or tentative when algorithm is unable to decide between two other categories. Two models chosen for evaluation included (1) only important attributes or (2) important and tentative ones.

For classification task, two algorithms were chosen: logistic regression (LR) and Support Vector Machine (SVM). Logistic regression is a standard in applied statistics when it comes to classification of individuals into two classes (Salazar et al., 2012). Studies have shown that it can often compete even with the most complex and novel methods such as artificial neural networks (Dreiseitl & Ohno-Machado, 2002). The logistic model is based on logistic function, the range of which is between 0 and 1. This range is the reason for model's popularity as it is able to describe probabilities, with values that are always between these values (Kleinbaum, 2013).

Support Vector Machine has established itself as an important alternative to logistic regression in recent years. Unlike regression, it relies on geometric, not statistical, properties of data and attempts to find the best separating hyperplane between two (or more) classes of outcomes. Unlike LR, which is linear, SVM allows for non-linear class separation with the use of kernels (Shmilovici, 2005). Among its advantages, when compared to LR, SVM requires fewer variables to achieve a better or equivalent misclassification rate (Salazar et al., 2012). Compared to LR, SVM is, however, less interpretable, which may be seen as a disadvantage in certain settings.

Motivated settings for selected methods

In the R implementation of logistic regression, there are just two customization options: P-value, for the results stage, and, in the prediction stage, probability cutoff used to choose classes. Here, P-value was left at the default level of 0.05. Cutoff was chosen based on value maximizing Area under the ROC Curve. This metric was chosen instead of accuracy, as accuracy is known to be unreliable for imbalanced datasets like this one (Bekkar et al., 2013).

For SVM, there are multiple tuning options. First of all, one needs to choose between classification and regression. In this binary scenario, classification was the method of choice. Secondly, type of kernel needs to be chosen. For situations like this one, the most common choice is radial, which is what informed the choice in this study (Sreenivasa, 2020). While other options were tested, they did not bring about significant improvement across the chosen metrics. Linear kernel uses sigmoid function, just as logistic regression, so their results are very similar. The two other tuning parameters are gamma and cost. Gamma affects the flexibility of the decision boundary with higher value meaning greater impact of the chosen features, hence a more flexible boundary (Al-Mejibli et al., 2020). Cost is a penalty for misclassification, which limits the boundary's flexibility. Gamma and cost were chosen by hyperparameter tuning with 10-fold cross validation to be 0.5 and 4 respectively across all tested models.

Results

Table 3 below outlines the results of feature selection under two previously discussed algorithms: F-Selector and Boruta.

Table 3 – Results of feature selection

	Selected variables	
	F-Selector	**Boruta**

economic situation vs last year	x – optimal subset	x
financial situation	x – optimal subset	x
safety in neighborhood	x – optimal subset	x
trust in community members	x	x
condition of roads	x	x
age	x – optimal subset	x
country	x	x
quality of medical services	x	x
quality of schools	x	x
occupation	x	x
years of education	x	x
# of children	x	x
human rights protection	x	
victim of crime last year	x	x
“those who govern this country are interested in what people think”	x	
freedom of political expression	x	Tentative
freedom of press	x	
relationship status	x	
interest in politics	x	
political knowledge of respondent	x	

religious services attendance	x	
government employee asked you for a bribe	x	
voted in presidential elections	x	
attention to the news	x	
sex	x	
attended a town meeting	x	
municipality size	x	
police officer asked you for a bribe	x	
protest participation	x	
# of children in the house		

*x = selected variable; x-optimal subset = determined as a part of important features subset in F-Selector model

It is clear from the table that variables that emerged as important in both models are economic situation when compared with last year, current financial situation, safety and trust in community, condition of roads, country of residence, quality of schools, medical services as well as occupation, education, number of children and crime victimization. Freedom of political expression was classified as important in one and tentative in the other model pointing to its potential importance. These were compared with the variables found to be significant in the LR estimations. Table 4 below lists the significant variables along with their baselines (variable levels they were compared to). The feature names in bold are the ones that overlap with variables found to be important by both used algorithms. We can see that, for the most part, they overlap. The LR below does not, however, include the number of children. At the same time, unlike feature selection models, it lists religious attendance, respondent’s sex and bribery as significant variables.

Table 4 – significant variables from tested models (LR)

	Boruta	**Boruta with tentative**	**F-Sel. optimal**	**F-Selector**
(Intercept)	1.82 ***	1.80 ***	1.14 ***	1.73 ***
	(0.29)	(0.29)	(0.16)	(0.35)
Country: El Salvador	-0.91 ***	-0.92 ***		-0.99 ***
(baseline: Mexico)	(0.18)	(0.18)		(0.19)
Honduras	-0.67 ***	-0.67 ***		-0.72 ***
	(0.19)	(0.19)		(0.20)
Nicaragua	-0.62 **	-0.62 **		-0.67 ***
	(0.19)	(0.19)		(0.20)
Ecuador	-0.36	-0.36		-0.39 *
	(0.19)	(0.19)		(0.19)
Peru	-0.37 *	-0.38 *		-0.40 *
	(0.17)	(0.17)		(0.18)
Brazil	-0.53 **	-0.54 **		-0.59 **
	(0.18)	(0.18)		(0.19)
Jamaica	-0.93 ***	-0.93 ***		-0.93 ***
	(0.19)	(0.19)		(0.20)
Financial status: Not enough, and are stretched	-0.40 **	-0.40 **	-0.52 ***	-0.43 **

(baseline: good enough and can save from it)	(0.15)	(0.15)	(0.14)	(0.15)
Financial status: Not enough, and having a hard time	-0.68 ***	-0.68 ***	-0.92 ***	-0.69 ***
	(0.16)	(0.16)	(0.15)	(0.16)
Interpersonal trust: Not Very Trustworthy	-0.23 *	-0.23 *		-0.21 *
(baseline: very trustworthy)	(0.10)	(0.10)		(0.10)
Interpersonal trust: Untrustworthy	-0.37 **	-0.37 **		-0.36 **
	(0.12)	(0.12)		(0.12)
age	-0.26 ***	-0.27 ***	-0.18 ***	-0.28 ***
	(0.05)	(0.05)	(0.03)	(0.05)
Occupation: Actively looking for a job	-0.22 *	-0.22 *		-0.20
(baseline: working)	(0.11)	(0.11)		(0.11)
Safety: Somewhat Unsafe	-0.53 ***	-0.53 ***	-0.60 ***	-0.54 ***
(baseline: very safe)	(0.11)	(0.11)	(0.10)	(0.11)
Safety: Very Unsafe	-0.45 ***	-0.45 ***	-0.63 ***	-0.50 ***
	(0.12)	(0.12)	(0.11)	(0.12)
Victim of crime in past year: No	0.23 **	0.22 **		0.24 **
	(0.08)	(0.08)		(0.08)
Quality of roads: Very Dissatisfied	-0.38 *	-0.37 *		-0.37 *

(baseline: very satisfied)	(0.18)	(0.18)		(0.19)
Quality of public schools: Very Dissatisfied	-0.39 *	-0.39 *		-0.37 *
(baseline: very satisfied)	(0.18)	(0.18)		(0.18)
Financial status vs last year: Worse	-0.63 ***	-0.63 ***	-0.72 ***	-0.61 ***
(baseline: better)	(0.11)	(0.11)	(0.10)	(0.11)
Religious attendance: Once or twice a year				-0.29 *
(baseline: more than once per week)				(0.13)
Sex: Female				0.19 *
				(0.08)
Police officer asked for a bribe (past year): Yes				0.23 *
				(0.12)
N	4008	4008	4008	4008
AIC	4997.43	5000.59	5116.59	5032.67
BIC	5274.45	5290.21	5179.55	5542.65
Pseudo R2	0.20	0.20	0.14	0.21

*** p < 0.001; ** p < 0.01; * p < 0.05.

The below results section summarizes some of the prediction results. The two metrics of interest for this study are accuracy and F-1. For each of the observations in test set, where the model is not given the life satisfaction class, the model predicts whether it belongs to category 1 (satisfied) or 0 (dissatisfied). The correct classifications, therefore, assign 1 where the dataset had a value 1 and 0 where it was a 0. Combinations of 0 in the place of 1 or 1 in the place of 0 are misclassified. Accuracy is a share of correct classifications. Hence, the highest value is 1 and the lowest 0. All the models below are in the 50-65% range, signifying moderate performance F-1 responds to the needs of imbalanced datasets, where the model could classify everything as one class and still achieve high accuracy. For example, if 90% of observations belong to class 1, 90% accuracy can be easily achieved by assigning 1 as all predictions, yet this would make the model very poor at differentiating between classes. F-1 is the harmonic mean of precision and recall. In other words, it looks at the combination of correct positive classifications (precision) and correctly classified actual positive values (recall) (*Machine Learning*, n.d.). In this way, it looks at accuracy in a more comprehensive manner. The values of F-1 are on the same scale as these of accuracy with higher value signifying better results (*Machine Learning*, n.d.). It can be seen that the values for estimated models are between 65 and just above 75%. Once again, this could be seen as moderate accuracy.

Table 5 – Prediction results from tested models

Logistic regression: prediction results	
Boruta	Accuracy: 0.6387

	F-1: 0.7511
Boruta with tentative	Accuracy: 0.6068 F-1: 0.7219
F-Selector optimal subset	Accuracy: 0.6233 F-1: 0.7404
F-Selector	Accuracy: 0.6211 F-1: 0.7355
Support vector machine: prediction results	
Boruta	Accuracy: 0.6061 F-1: 0.7258
Boruta with tentative	Accuracy: 0.6127 F-1: 0.7303
F-Selector optimal subset	Accuracy: 0.6413 F-1: 0.7577
F-Selector	Accuracy: 0.5346 F-1: 0.6543

Conclusion and Discussion

Both LR and SVM give poor to moderate performance at predicting subjective wellbeing. All models had similar predictive performance despite differences between number of variables and used algorithms. The slight differences across accuracy metrics in each algorithm were not consistent between them i.e. there is no consistent performance rank for the tested models. In other words, the performance rank is different when we use the same models for LR and SVM. Therefore, little can be said about the relative importance of some features over others from the prediction results alone. The low to moderate accuracy, despite the use of reliable and widely accepted methods, suggests that the used variables do not include all that contribute to wellbeing. Hence, more research is needed to investigate these. While one could approach the task from a data-scientific perspective and try out all possible combinations of the 535 available variables, this method is computationally expensive and ethically questionable. It could lead, even without it being directly stated, to beliefs about non-existent causal links, which could potentially lead to harmful or, at best, suboptimal outcomes. Therefore, literature-based variable choice was the method used in this paper.

The two feature selection models, Boruta and F-Selector, are in agreement on the importance of financial situation, safety and trust in community, condition of roads, country of residence, quality of schools, medical services as well as occupation, education, number of children and crime victimization. The results of logistic regression, specifically significance of variables, confirm all of these except for the number of children. Religiosity, sex and bribery appear as significantly related to the outcome despite not being part of the variable selection models' overlap. The high overlap between the two algorithms confirms relative importance of certain variables over others. By confirming, to a great extent, the previous research, this paper brings more clarity on the topic. The mixed results e.g. with respect to religiosity or bribery, point to the need of more research on the topic.

This research paper finds no support for the literature stating that financial factors may be less important in Latin America. On the contrary, the results suggest that financial comfort may be one of the most important predictors of wellbeing. Similarly, it finds that social development in one’s region is important to one’s wellbeing as are safety and certain demographic characteristics. The fact that some of these variables are in scope of action of local municipalities points to potential policy implications. There is not enough evidence to confirm the importance of religiosity in the region, yet this variable emerged as significant in Logistic Regression, hence more research could bring clarity on the topic. The study of wellbeing has important implications for societies at large. This paper, by using novel methods to confirm some previous studies and put others in question, opens up the field for further research on the topic. In later research, other algorithms could be used to predict wellbeing on the same data, or they could use other pre-processing methods to see whether this changes the results in a significant manner.

References

Al-Mejibli, I. S., Alwan, J. K., & Abd, D. H. (2020). The effect of gamma value on support vector machine performance with different kernels. International Journal of Electrical and Computer Engineering (IJECE), 10(5), 5497. https://doi.org/10.11591/ijece.v10i5.pp5497-5506

Bekkar, M., Djema, H., & Alitouche, T. A. (2013). Evaluation measures for models assessment over imbalanced data sets. Journal of Information Engineering and Applications, 3, 27–38.

Bhaskaran, K., & Smeeth, L. (2014). What is the difference between missing completely at random and missing at random? International Journal of Epidemiology, 43(4), 1336–1339. https://doi.org/10.1093/ije/dyu080

Conci, P. (2017, April 19). Why are Latin Americans happier than their GDP would suggest? Ideas Matter. https://blogs.iadb.org/ideas-matter/en/latin-americans-happier-gdp-suggest/

Corral, M. (2011). The Economics of Happiness in the Americas. AmericasBarometer Insights, 58.

Data sets. (n.d.). LAPOP. Retrieved July 20, 2022, from https://www.vanderbilt.edu/lapop/raw-data.php

Diego-Rosell, P., Tortora, R., & Bird, J. (2016). International determinants of subjective well-being: Living in a subjectively material world. Journal of Happiness Studies, 19(1), 123–143. https://doi.org/10.1007/s10902-016-9812-3

Dolan, P., Peasgood, T., & White, M. (2008). Do we really know what makes us happy? A review of the economic literature on the factors associated with subjective well-being. Journal of Economic Psychology, 29(1), 94–122. https://doi.org/10.1016/j.joep.2007.09.001

Dreiseitl, S., & Ohno-Machado, L. (2002). Logistic regression and artificial neural network classification models: A methodology review. Journal of Biomedical Informatics, 35(5–6), 352–359. https://doi.org/10.1016/s1532-0464(03)00034-0

Dulam, T., Grift, Y., & Van den Berg, A. (2021). Economic hardship, institutions and subjective well-being in Latin America. U.S.E. Research Institute Working Paper Series 21-06.

Garzon, P., & Ruprah, I. (2015). Religion as an Unemployment Insurance and the Basis of Support for Public Safety Nets: The Case of Latin America and the Caribbean. IDB Working Paper, IDB-WP-601.

Kleinbaum, D. G. (2013). Logistic regression: A self-learning text. Springer Science & Business Media.

Kursa, M. B., & Rudnicki, W. R. (2010). Feature Selection with theBorutaPackage. Journal of Statistical Software, 36(11). https://doi.org/10.18637/jss.v036.i11

Lee, W., & Seo, K. (2022). Downsampling for binary classification with a highly imbalanced dataset using active learning. Big Data Research, 28, 100314. https://doi.org/10.1016/j.bdr.2022.100314

Machine learning. (n.d.). Google Developers. Retrieved July 19, 2022, from https://developers.google.com/machine-learning/crash-course/classification/precision-and-recall

Near, J. P., Smith, C. A., Rice, R. W., & Hunt, R. G. (1984). A comparison of work and nonwork predictors of life satisfaction. Academy of Management Journal, 27(1), 184–190. https://doi.org/10.5465/255966

Oishi, S., Diener, E., Lucas, R. E., & Suh, E. M. (2009). Cross-Cultural variations in predictors of life satisfaction: Perspectives from needs and values. In Social Indicators Research Series (pp. 109–127). Springer Netherlands. http://dx.doi.org/10.1007/978-90-481-2352-0_6

Romanski, P., Kotthoff, L., & Schratz, P. (2021, February 16). CRAN. Package FSelector. https://cran.r-project.org/web/packages/FSelector/index.html

Ruggeri, K., Garcia-Garzon, E., Maguire, Á., Matz, S., & Huppert, F. A. (2020). Well-being is more than happiness and life satisfaction: A multidimensional analysis of 21 countries. Health and Quality of Life Outcomes, 18(1). https://doi.org/10.1186/s12955-020-01423-y

Salazar, D., Vélez, J., & Salazar Uribe, J. (2012). Comparison between SVM and Logistic Regression: Which One is Better to Discriminate? Revista Colombiana de Estadística, 35, 223–237.

Shmilovici, A. (2005). Support Vector Machines. In O. Maimon & L. Rokach (Eds.), Data Mining and Knowledge Discovery Handbook (pp. 257–276). Springer US. https://doi.org/10.1007/0-387-25465-X_12

Sreenivasa, S. (2020, October 12). Radial basis function (RBF) kernel: The go-to kernel. Towards Data Science. https://towardsdatascience.com/radial-basis-function-rbf-kernel-the-go-to-kernel-acf0d22c798a

Stone, L. (2022, June 29). After COVID: Unhappiness is worse among single and non-religious Americans. Institute for Family Studies. https://ifstudies.org/blog/after-covid-unhappiness-is-worse-among-single-and-non-religious-americans

Tov, W., & Au, E. W. M. (2013). Comparing well-being across nations: Conceptual and empirical issues. Oxford University Press. http://dx.doi.org/10.1093/oxfordhb/9780199557257.013.0035

Valente, R. R., & Berry, B. J. L. (2016). Dissatisfaction with city life? Latin America revisited. Cities, 50, 62–67. https://doi.org/10.1016/j.cities.2015.08.008

Van Buuren, S. (2018). Flexible Imputation of Missing Data, Second Edition. CRC Press.

Vanderbilt University. (2017). Americas Barometer. https://www.vanderbilt.edu/lapop/raw-data.php

Zhang, X. (2017). Support Vector Machines (C. Sammut & G. I. Webb, Eds.; pp. 1214–1220). Springer US. https://doi.org/10.1007/978-1-4899-7687-1_810