Anotações - Religious Affiliation in the Twenty-First Century

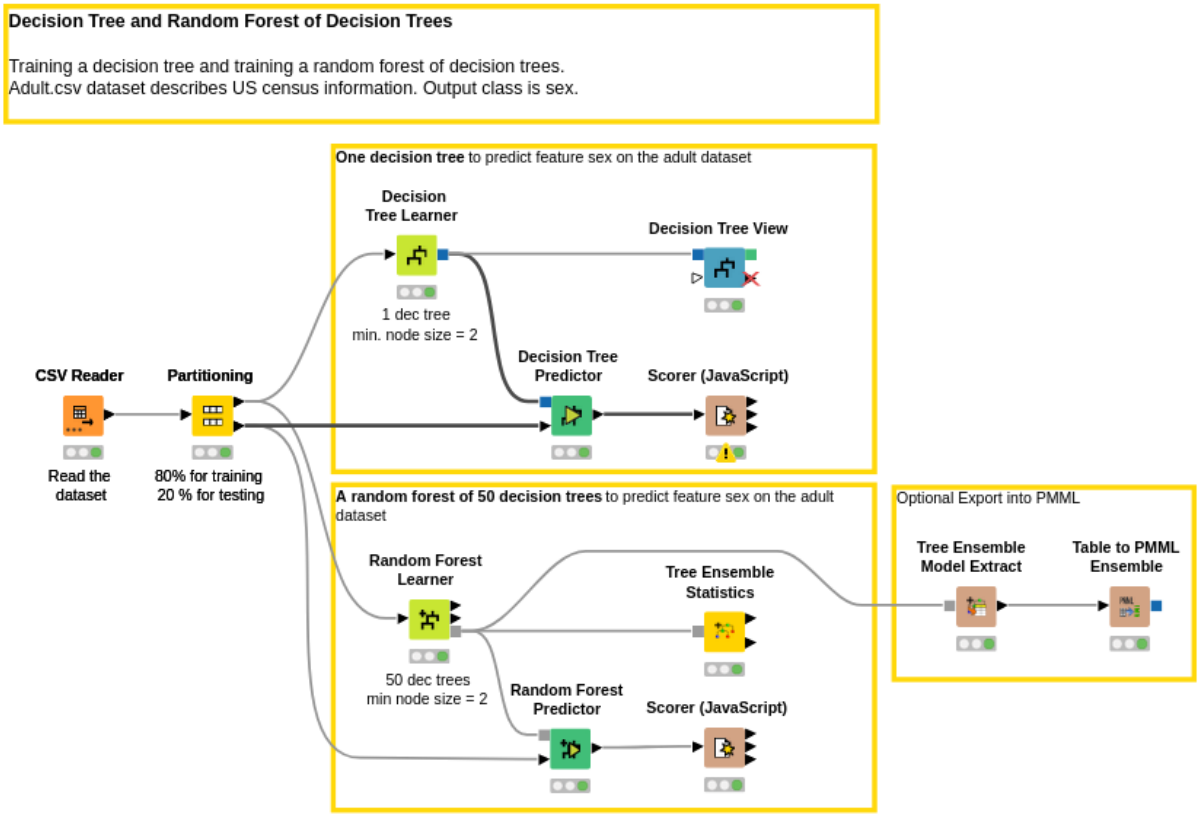
Religious Affiliation in the Twenty-First Century: A Machine Learning Perspective on the World Value Survey

Abstract

Análise quantitativa dos dados coletados globalmente pelo World Value Survey.
Os dados são usados para estudar as tragetórias de mudança das crenças religiosas, valores e comportamentos dos individuos nas sociedades.

Utilizando Random Forest, pretendemos identificar os factores chave de religiosidade e classificar os inquiridos como religiosos ou não-religiosos.

Random forest - knime hub



https://stats.stackexchange.com/questions/172842/best-practices-with-data-wrangling-before-running-random-forest-predictions

Utilizamos tecnicas de resampling para balancear os dados e melhorar "imabalanced learning performance metrics".

Ressampling - knime hub
Resampling in Supervised Fraud Detection Models

> The results of the variable importance analysis suggest that Age and Income are the most important variables in the majority of countries.

Os resultados são discutidos com teorias sociológicas fundamentais em relação a religião e comportamento humano.
Este estudo é uma aplicação de machine learning em identificar os padrões subjacentes nos dados de 30 países que participam no survey.

Introduction

> World Values Survey (WVS) is a non-commercial, non-governmental program dedicated to education and research. WVS is a publicly available dataset from a global survey since 1981. WVS aims to assess the population of the participating countries on all continents, on different criteria, and to **identify what people most value in life, their impact on cultural stability, social and political issues, and their development over time.**

Identificar o que as pessoas mais valorizam na vida, o seu impacto na estabilidade cultural, social, e em problemas políticos, e o seu desenvolvimento ao longo do tempo.

The data consists of 7 rounds of surveys between 1981 and 2022.

Neste estudo pretendemos utilizar machine learning para explorar os dados e responder duas questões:

1. What are the strong indicators of religious beliefs and practice of such doctrines among the population?
2. Can we use machine learning to classify people as religious or non-religious using historical data?

3. Quais são os indicadores fortes de crenças religiosas e da prática dessas doutrinas entre a população?
4. Podemos utilizar machine learning para classificar pessoas como religiosas ou não religiosas utilizando dados históricos?

> Machine learning is an **umbrella term** for a group of algorithms used to identify the underlying patterns present in the data in an iterative process [2]. Machine learning involves statistical analysis, regression, studying the correlations between variables and their impact on the dependent variable, feature important analysis and feature selection, methods to address data-related issues such as resampling, and algorithms like random forest for classification.

Secularization Theory

Secularization theory is a sociological doctrine that claims religion, in its traditional sense, is in a terminal decline as the world moves towards industrialization and economic growth.

> The proponents of **secularization theory** argue that scientific advances, urbanization, and mass education decrease the influence of religion in societies.

Preferia que não fosse verdade, então vou ver se dá para provar isso

Rational Choice Theory

> The **rational choice theory** states that individuals' decisions are based on calculations to optimize their outcomes, utility, and satisfaction. Based on this theory, personal interests and goals are the main drives of human decision-making. Rational choice theory is an influential perspective that has been applied to a wide range of social and economic phenomena. The supporters of the rational choice theory believe that religion follows the same principles of the free market and individuals decide on their level of commitment to religion while considering the costs and benefits, and that is why religious entities still exist in the 21st century (churches, mosques, temples, etc.)

Decisões de indivíduos são baseadas em cálculos para otimizar resultados, utilidade e satisfação. Tendo em conta esta teoria, interesses pessoais e objetivos são os principais motores de decisão humana => Indivisuals decide on their level of commitment to religion while considering the costs and benefits.

Existential Security Theory

> Drawing from human's desire for stability and security, existential security theory studies the environmental, social, and economic threats to human security. Such threats drive societies towards collective actions that affect individuals and societies in positive or negative ways. Positive impacts are forming communities that lead to growth, development, and an increase in the overall well-being of society. Negative impacts are tendencies towards extreme ideologies, conflict, and sometimes violence. Focused on the patterns of change in social and political structures, this theory also argues that lack of financial security causes religion to be more present in an individual's life

Estuda as ameaças ambientais, sociais e económicas para a segurança humana. Tais ameaças levam sociedades para ações coletiva que afetam indivíduos e sociedades em maneiras positivas e negaticas.
Impactos positivos são a formação de comunidades que levam a crescimento, desenvolvimento e um aumento de bem-estar da sociedade.
Impactos negativos são tendências para ideologias extremas, conflitos e por vezes violência.
Focando nos padrões de mudança social e estruturas políticas, esta teoria também argumenta que a falta de segurança financeira causa uma maior presença de religiao.

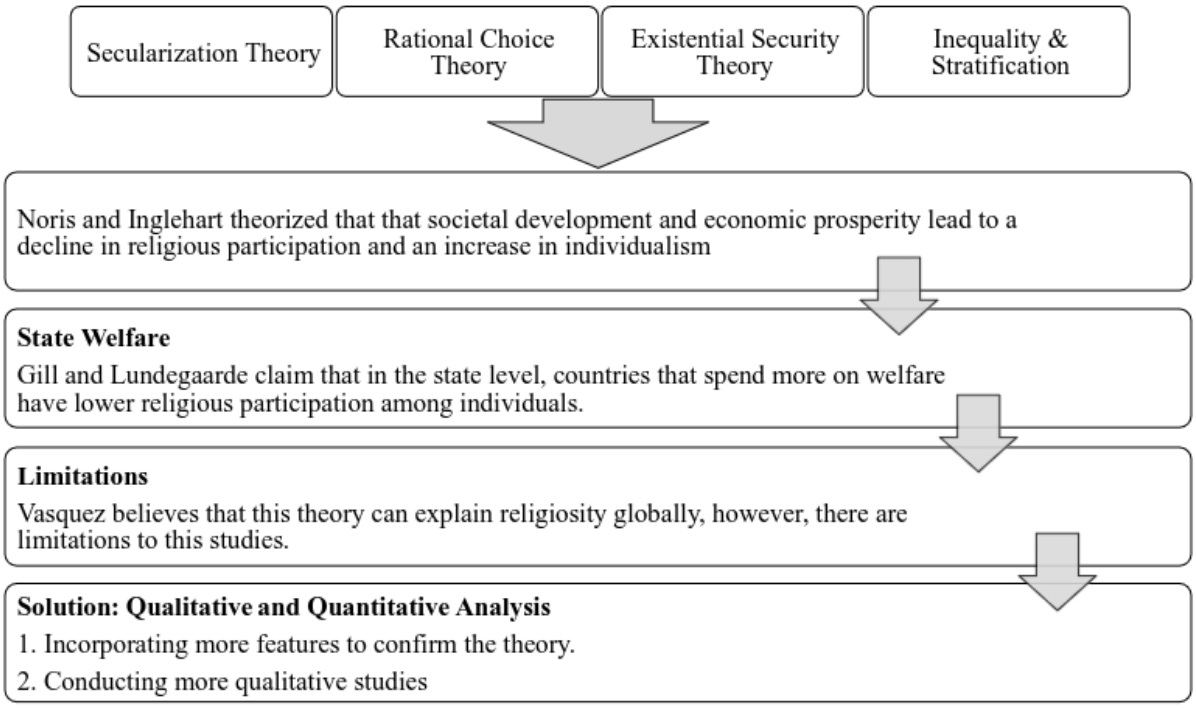
Social Stratification

Social stratification is the inherent non-uniform distribution of wealth, resources, power, and social status within societies. Inequality results from unequal stratification of resources based on class, race, and gender that affects income, education, health, etc. Over the years, numerous studies have investigated the environmental factors shaping personal values and religious beliefs. However, critiques such as Vasquez 20, 21 argue that more qualitative

and quantitative analysis of nationally representative data is required to verify the theory on global data. Therefore, this study quantitatively analyzes WVS data using data-driven techniques to support the theories talked about.

Outras teorias não faladas no paper

- **Functionalism**: religion provides social cohesion, meaning, and purpose for individuals and communities. According to functionalism, religion helps to maintain social order and stability by providing moral guidance and a sense of belonging.
- **Symbolic Interactionism**: this perspective focuses on the ways in which individuals and groups interpret and construct meanings through their interactions. In the context of religion, symbolic interactionism emphasizes how religious symbols, rituals, and practices shape individuals' identities and social interactions. It recognizes the importance of religious experiences and meanings in people's lives.
- **Archetypal Theory**: Jung proposed the concept of archetypes, which are universal symbols and patterns that exist in the collective unconscious of humanity. These archetypes manifest in myths, symbols, and religious imagery across cultures and historical periods. In the context of religion, Jungian scholars explore how religious stories, rituals, and symbols reflect archetypal themes such as the hero's journey, the divine mother, the trickster, and the wise old man. They suggest that these archetypes resonate with deep aspects of the human psyche and contribute to the meaning-making process in religious experiences.
- **Individuation and Spiritual Development**: Jungian psychology emphasizes the process of individuation, which involves integrating unconscious aspects of the psyche into conscious awareness to achieve psychological wholeness and self-realization. In the context of spirituality and religion, Jungian theorists explore how religious beliefs and practices can facilitate individuation by providing symbolic frameworks for personal growth, transformation, and integration of the psyche. They view spiritual development as a journey of self-discovery and inner exploration that may intersect with religious traditions and experiences.
- **Cultural Theory**
- **Postmodernism** (sim, tbm isto)



Analysis

Overview of the Data

The dependent variable, religiosity, is constructed on the basis of three items related to religion; **salience, identification and behaviour,** as suggested by Welzel.
Calculating religiosity:

1. How important is religion in your everyday life? -- salience
2. Are you a religious person? -- identification
3. Apart from weddings and funerals, how often do you attend religous services? -- behaviour

> The response values are scaled and combined to get a continuous value between 0 and 1 as the religiosity score, while 0 means Not religious, and 1 is Highly religious. In this study, the variables are converted to 0 and 1 for binary classification.

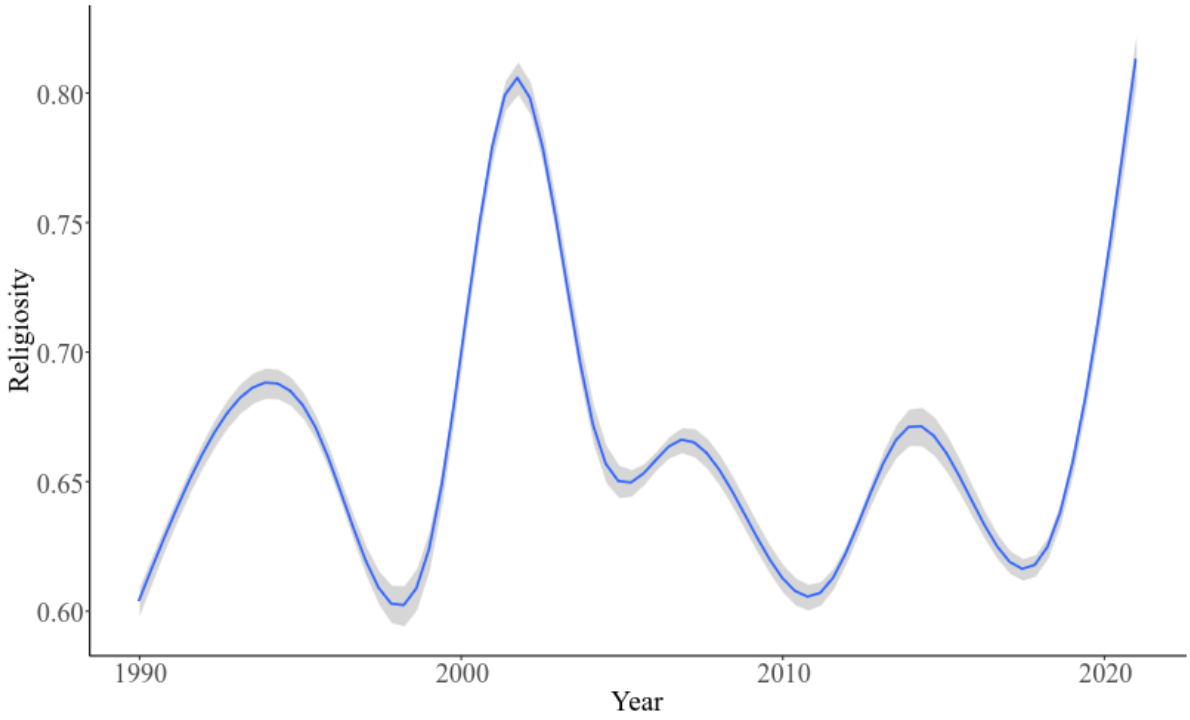
Idependent categorical variables in the data:

Table 1: Independent variables

Variable	Levels
Education	Pre-primary Elementary Secondary Tertiary
Life Satisfaction	Very satisfied Satisfied Neutral Unsatisfied Very unsatisfied
Political View	Left Left-leaning Moderate Right-leaning Right
Social Class	Lower class Working class Middle class Higher class

> An important variable in understanding religiosity is the **geographical** region where the participants reside. The underlying cultural, geopolitical, and historical characteristics of the countries impact religiosity. Studying the complete list of countries present in the survey is beyond the scope of the paper. Therefore, we have limited the study to 30 countries with an adequate sample size for the classification task. The countries are selected from different regions with various cultural, political, and socio-economic structures.

Religiosity changes over time (1995-2022)



We used the random forest model to **evaluate and rank the independent variables by their importance in predicting the religiosity score and classifying individuals based on their scores.**

> Random forest is used for classification and regression in different domains. A random forest is a collection of hierarchical structures with a set of rules for dividing large, diverse data into smaller homogeneous groups with respect to a target variable. Random forest uses decision trees to reduce error by constructing a multitude of decision trees in parallel during training and picking random splits in the trees. It outputs a prediction that is the mean or the majority vote of the individual trees. Random forest is a powerful method with interpretable results that can handle missing values and noisy data. Therefore, it is selected as the method of choice in this study. The splitting criterion in the decision trees is the measure of impurity, which is the measure of the homogeneity of the target variable. Node impurity is the average decrease in the impurity as a result of splitting the data based on the values of the variable over all the trees in the forest. The variables are ranked from the highest increase in node impurity to the lowest, the highest being the most important. The training continues until the impurity at each node is minimized and the highest accuracy is achieved. In this study, we use the metric Increase in Node Purity calculated in R.

Table 2: Summary statistics of religiosity in the selected countries

Country	Count	Mean	Standard Deviation	Median
Nigeria	7359	0.93	0.13	1
Ethiopia	2610	0.89	0.17	0.94
Kenya	1224	0.88	0.16	0.94
Zimbabwe	3530	0.88	0.18	0.94
Bangladesh	3873	0.88	0.15	0.94
Indonesia	5517	0.88	0.15	0.94
Rwanda	2933	0.84	0.16	0.89
Yemen	916	0.81	0.18	0.78
South Africa	10864	0.78	0.24	0.89
Iran	5736	0.78	0.21	0.83
India	11090	0.77	0.21	0.83
Brazil	7229	0.75	0.24	0.83
Peru	6167	0.74	0.22	0.78
Iraq	6470	0.72	0.21	0.67
Saudi Arabia	1227	0.71	0.22	0.72
Italy	645	0.71	0.25	0.78
Turkey	10259	0.71	0.25	0.78
Mexico	8822	0.70	0.27	0.78
Venezuela	3298	0.67	0.26	0.72
Kyrgyzstan	3607	0.66	0.24	0.67
United States	9845	0.63	0.33	0.72
Switzerland	1951	0.49	0.32	0.56
Spain	4855	0.46	0.34	0.50
Russia	8206	0.46	0.29	0.56

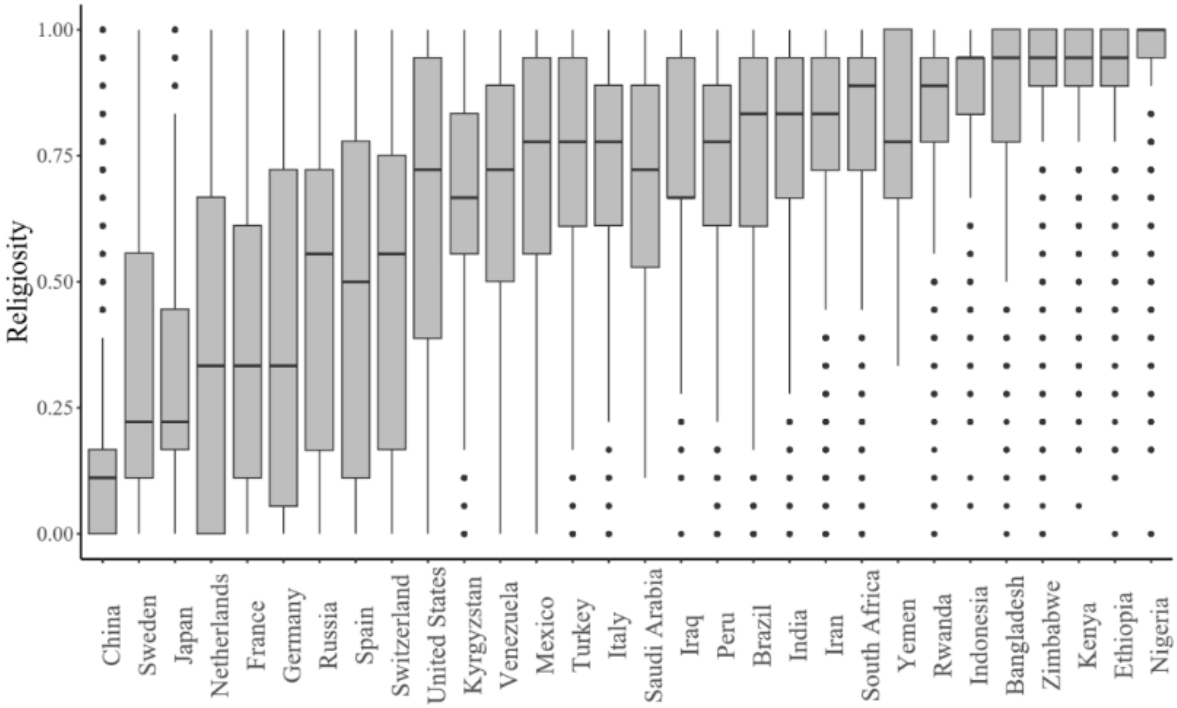


Fig. 4: Distribution of religiosity in the selected countries

Geographical background plays a critical role in the dispersion of religious beliefs and values.

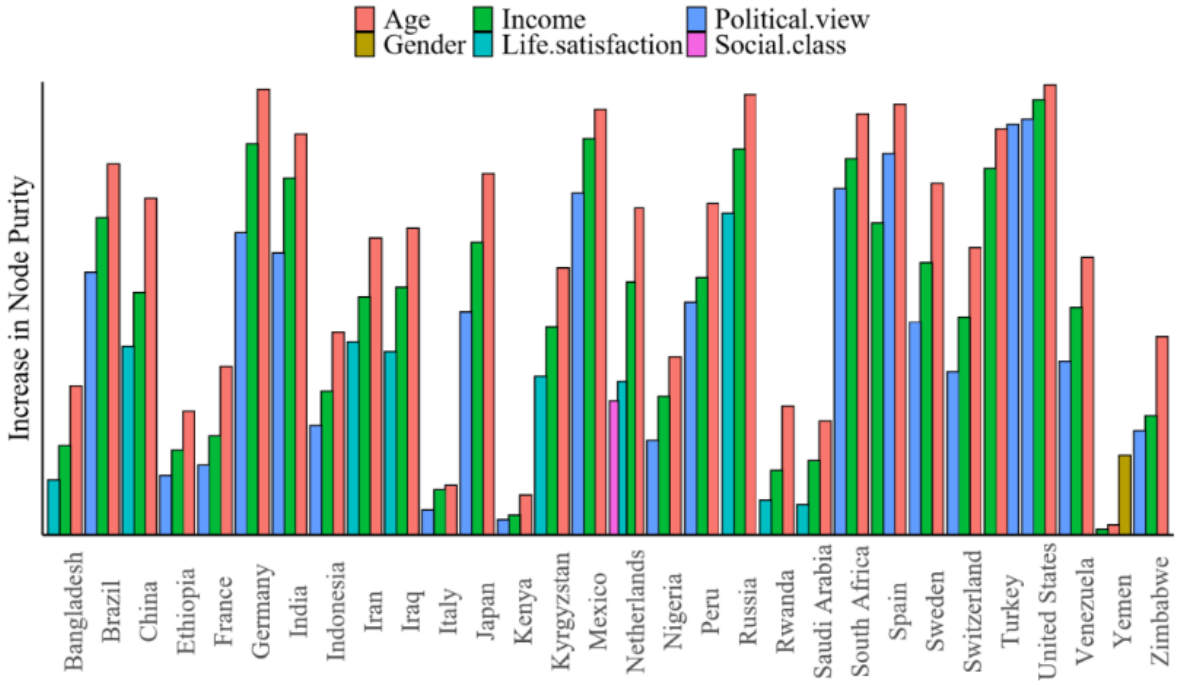


Fig. 5: Top three variables with the highest increase in node purity for each country

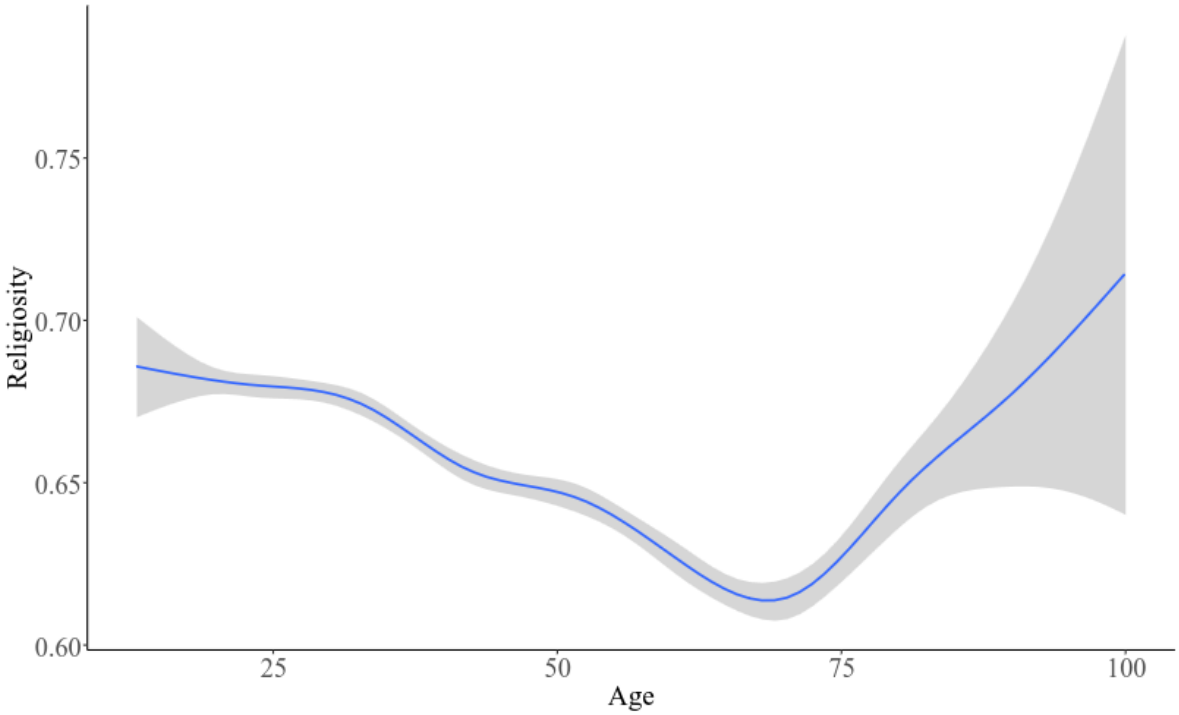
Age and Income are the top two most important variables in 93.3% of the countries.

The next most important variable is Political status in 60% of the countries and Life satisfaction in 26% of the countries.

Age

2.2.1 Age

The fitted curve in Figure 6 shows the relationship between religiosity and age in the data, and Figure 7 presents the distribution of religiosity by country.



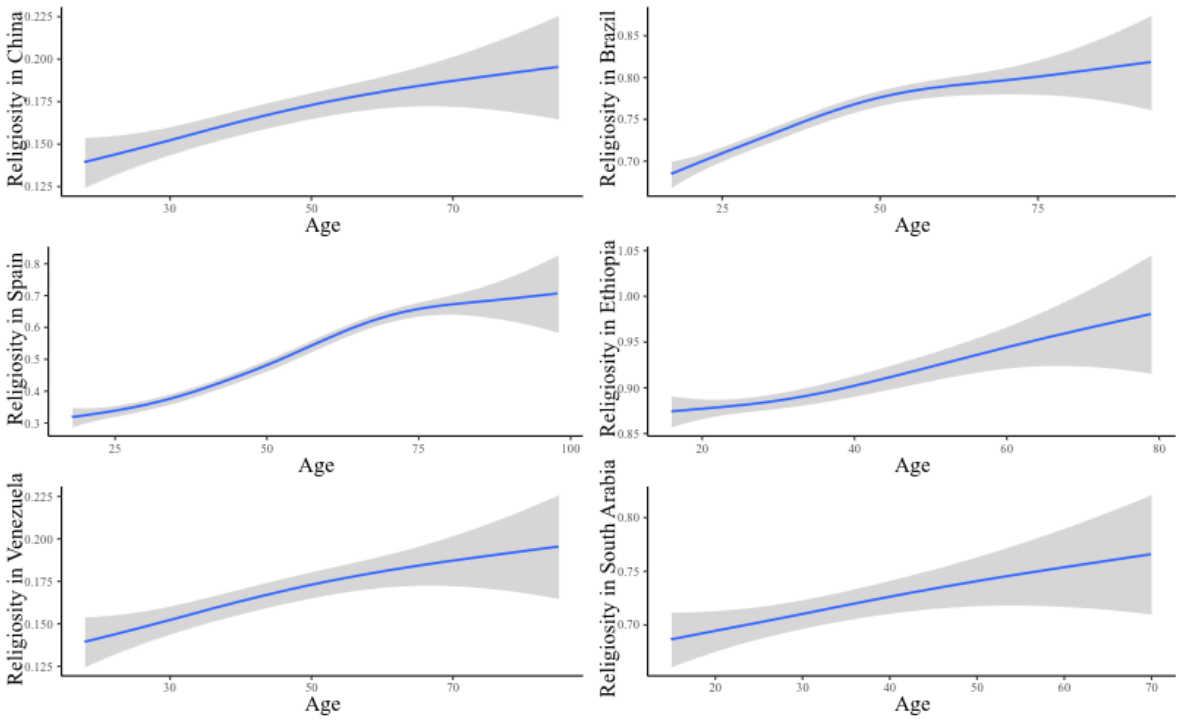


Fig. 8: The fitted curves show the increase in religiosity with age in participants from a sample of 6 countries present in the study. More details on the correlation between religiosity and age are provided in Table 3

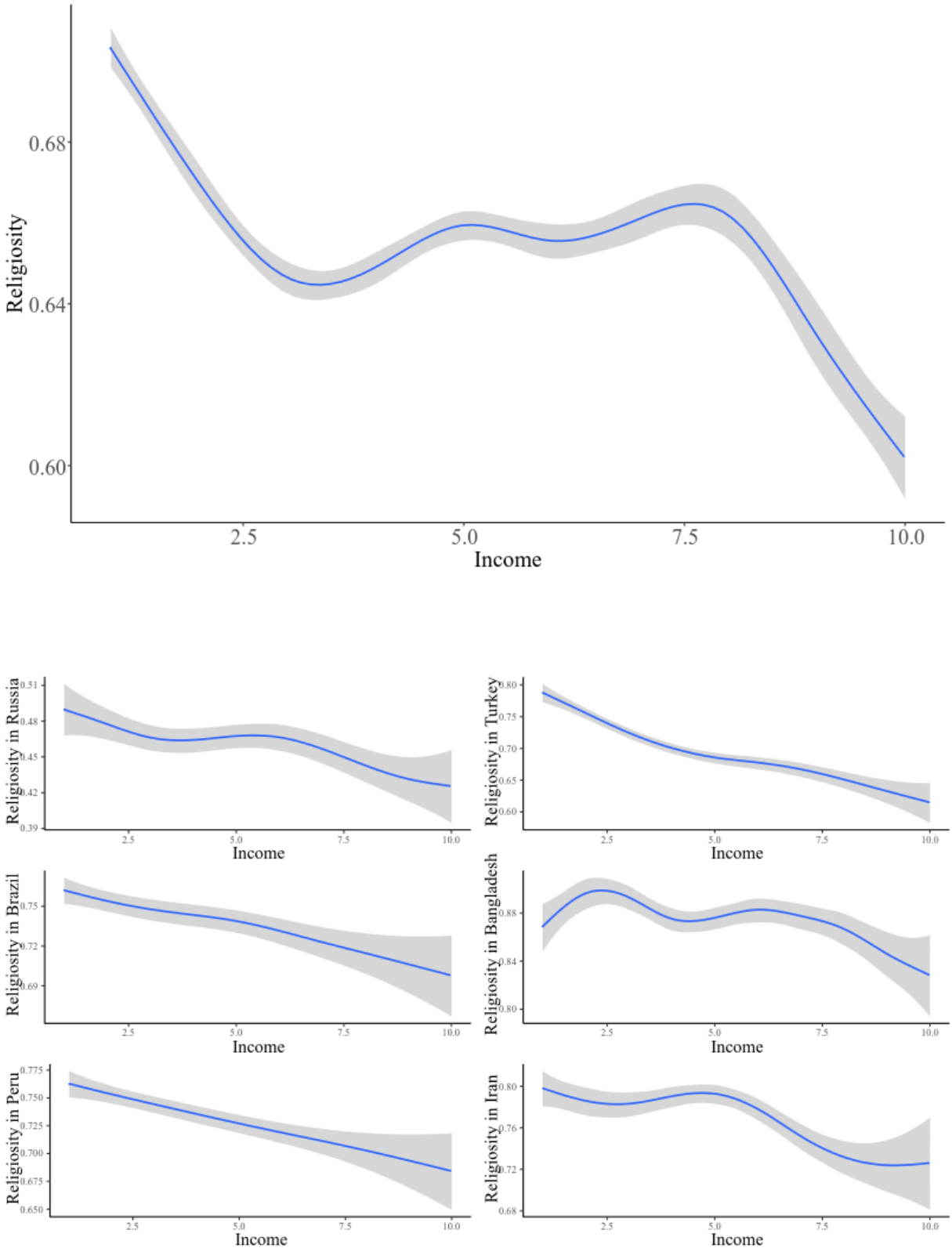
> Sociologists explain this behavior by existential security theory. Generally, people tend to experience more insecurity and unpredictability in their lives as they grow older. The challenges and insecurities they face as a result of aging drive people to rely on religion as a source of comfort and hope. Also, psychological studies support the idea that participating in religious services creates a sense of community which improves the mental and physical health of the elderly.

In conclusion, the results from our analysis are consistent with the idea that a decrease in existential security as a result of aging impacts life choices and perspective. While these insecurities are caused by the decline in physical, cognitive, and mental abilities, health problems, financial instability, and loss of friends and loved ones, it results in an increase in religiosity with age.

Income

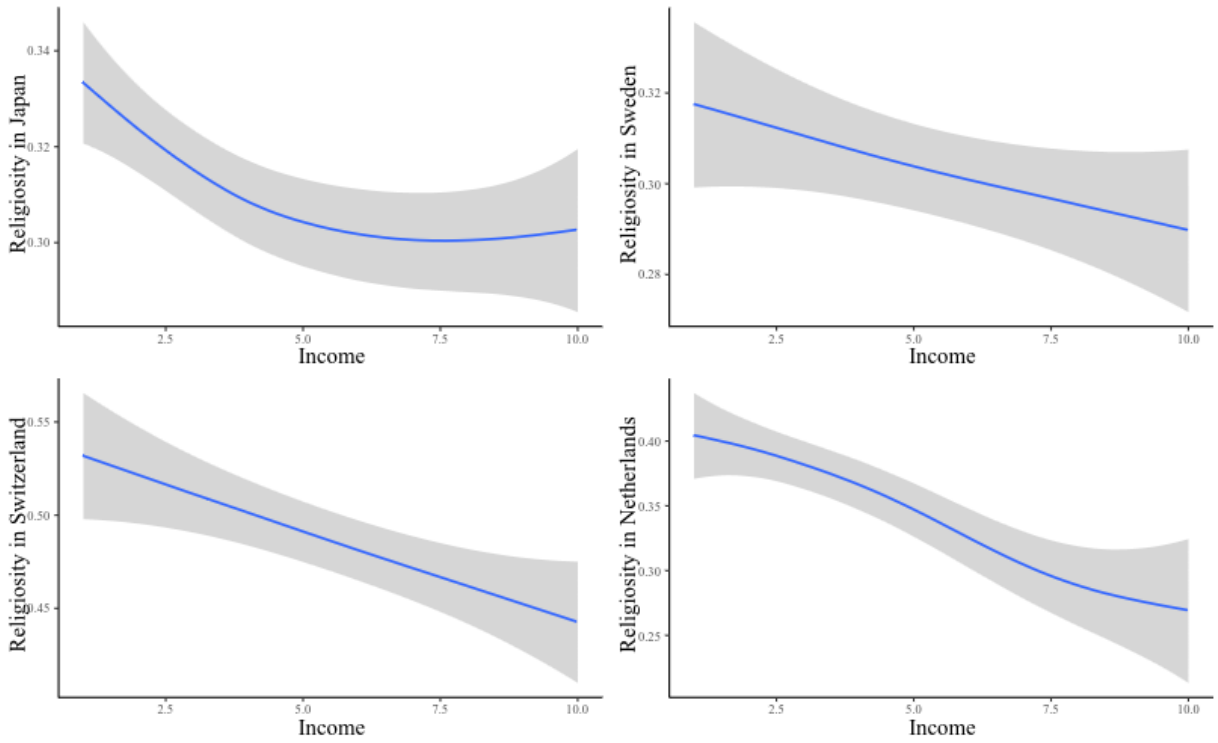
> The figure bellow resents the negative correlation between religiosity and income. Religiosity is often higher among people with lower incomes and declines as income levels grow. While this negative correlation is true for

> many countries, there are also exceptions to this. Therefore, we will have a closer look at the country-level data.



The decline in religiosity as a result of an increase in income levels in developing and underdeveloped countries is explained by the existential security theory caused by stratification and inequality in societies.

The secularization theory is another facet of the negative correlation between religiosity and income. While individuals are less likely to experience the insecurities caused by famine, disease, war, and environmental disasters, in developed counties,
industrialization, and scientific growth replace religious beliefs. Therefore the overall
religiosity levels decline. To support this idea:



Outlier:

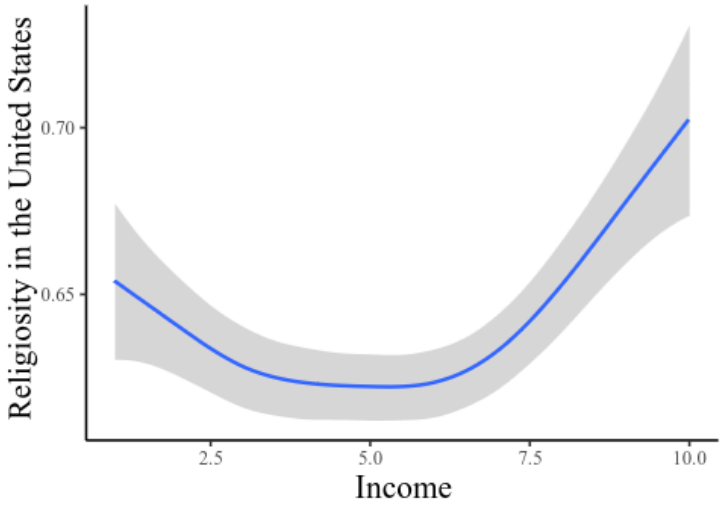


Fig. 12: Unlike many developed countries, religiosity and income are not always negatively correlated in the United States. In higher income levels, religiosity, and income are positively correlated

Em yemem (islamic country) onde religiao está profundamente interlaçada com
poder e riqueza.

> In Iran, religiosity is a marker of power and the wealthy elite.

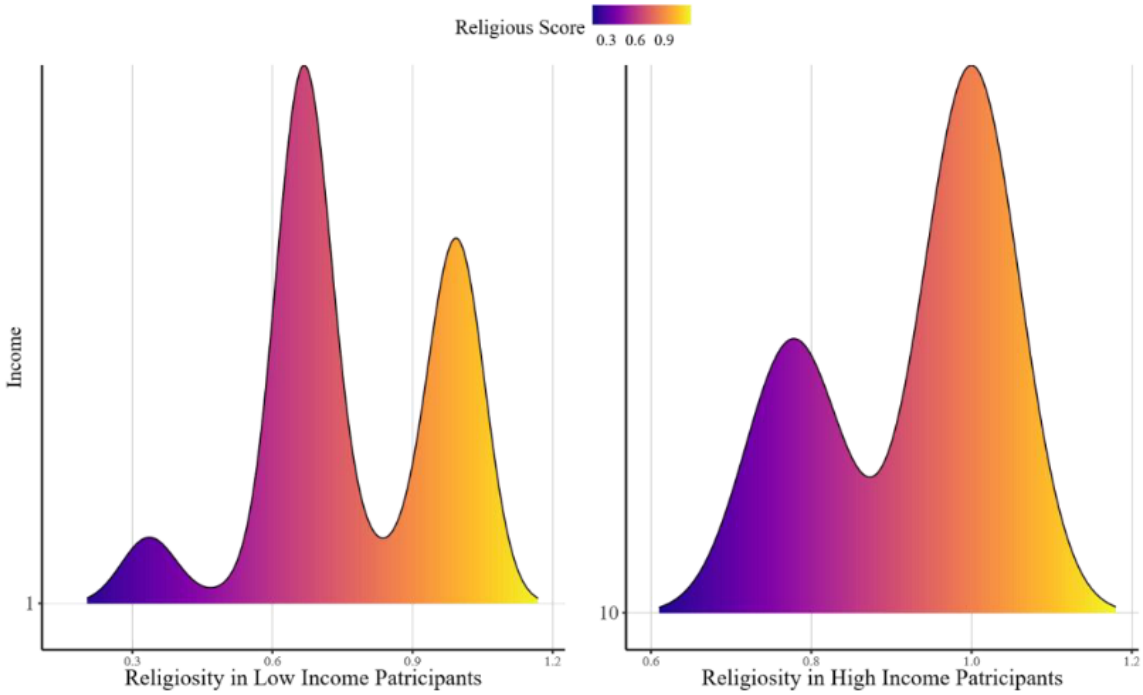


Fig. 13: The majority of individuals with high-income levels have higher religiosity scores. Also, the population of high-income, highly religious individuals is significantly larger than the population of individuals with high religiosity scores among low-income populations

> This religious culture among the wealthy and powerful in Yemen and Iran is in line with the rational choice theory, where personal goals and satisfaction play an important role in the decision to participate in religious gatherings.

Overall, the relationship between religion and income is very complex, and it is deeply influenced by the social and political structures in society. The cultural background, the level of industrialization, equality, and equity are also essential
contributors to religiosity.

Political Views

Referring to the secularization theory, scientific advances and growth in technology have reduced the influence of religion on societies. This has caused a shift in political views and a digression from conservatism.

Secularization theory suggests that people with liberal or progressive political views, such as supporters of individual rights, social equity, and environmental issues, mostly identify as less religious or non-religious. Less religious individuals tend to view the world through a secular lens. They prioritize rational and evidence-based thinking over traditional religious beliefs and values.

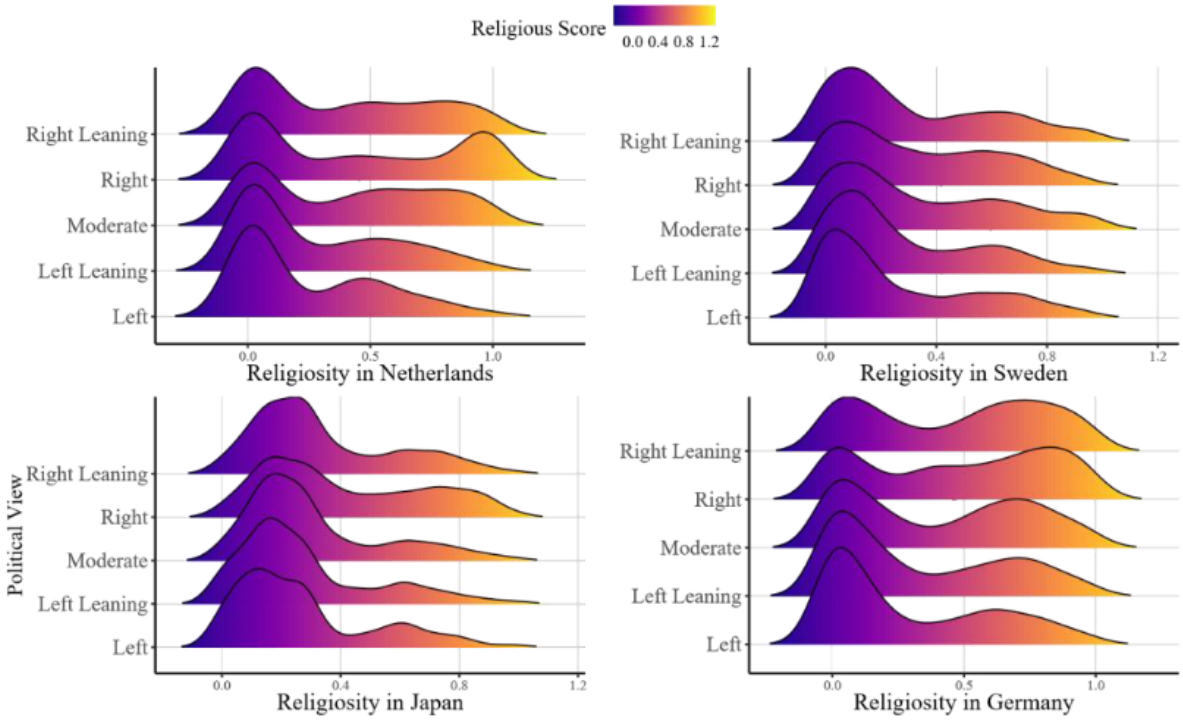


Fig. 15: Distribution of religiosity by political views in developed countries

The positive and statistically significant coefficients suggest that religiosity increases as political views become more conservative. However, the correlation between religiosity and political views is weak. The small value of adjusted R2 suggests the non-linear relationship between political views and religiosity in these countries.

Table 4: Correlation analysis and linear regression results

Country	Correlation Coef.	Adjusted R^2	P-value	F-statistic
Netherlands	0.129	0.016	5.86e-09	34.18
Sweden	0.058	0.003	0.00	12.82
Japan	0.108	0.011	$5.267e-13$	52.42
Germany	0.156	0.024	$< 2.2e-16$	140

The critics of the secularization theory argue that there are various counterexamples, and this theory doesn’t explain the resurgence of religious movements in some regions.

Life Satisfaction

The figure bellow presents the distribution of religiosity in different levels of life satisfaction in countries where life satisfaction is the third most important variable. These countries are Bangladesh, China, Iran, Iraq, Kyrgyzstan, Russia, Rwanda, and Saudi Arabia, which are mainly developing and under-developed countries.

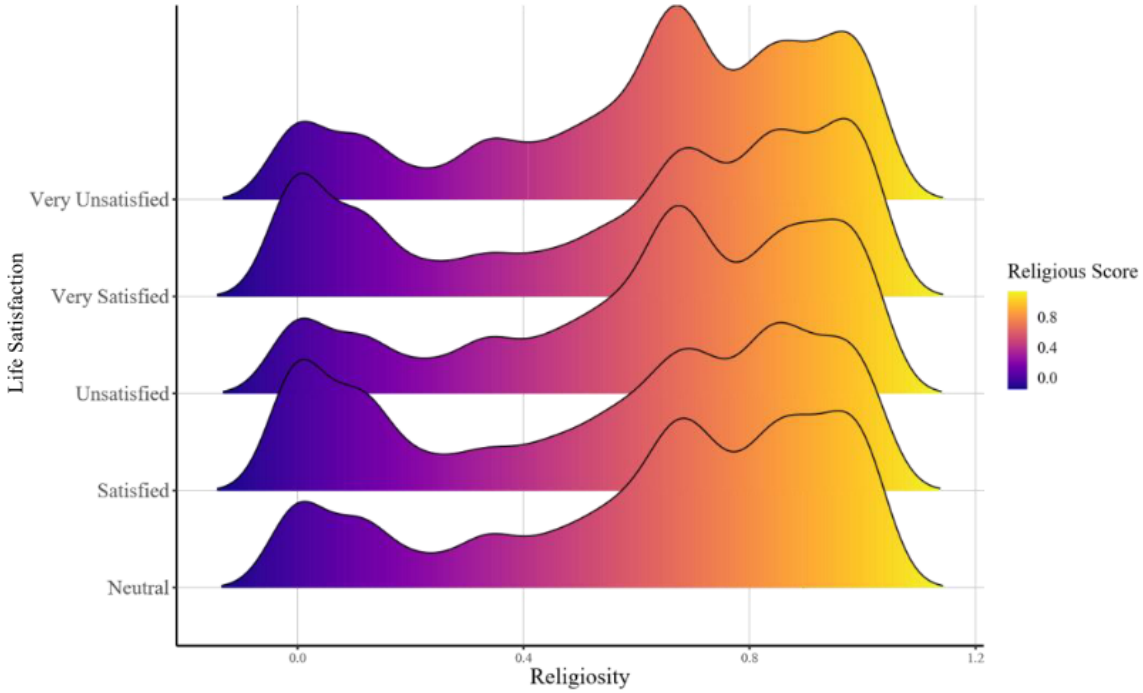


Fig. 17: Countries where life satisfaction is ranked as the third most important variable

Nonetheless, to better understand the relationship between religiosity and life satisfaction, we need to consider the influence of factors such as the type of religion or spirituality in question, the cultural context, and the individual’s personal beliefs and experiences.

Classification

Nesta secção, analisamos as relações entre várias "predictor variables" e religiosidade para treinar um modelo de machine-learning para classificar individuos como religiosos ou não-religiosos usando "historical data".

> By training machine learning models on historical data such as survey responses, we explore the patterns within the data and classify respondents based on their religious affiliations. We studied the key predictors and identified the factors that contribute to religiosity.

Cross validation:

- https://hub.knime.com/knime/spaces/Examples/04_Analytics/15_H2O_Machine_Learning/04_H2O_Crossvalidation~zzTdxkw7hfziD4Ik/current-state
- https://hub.knime.com/knime/spaces/Educators%20Alliance/Guide%20to%20Intelligent%20Data%20Science/Example%20Workflows/Chapter5/01_CrossValidation_Scorer_ROC~-0nN9BzUOCI6vCXl/most-recent
- https://hub.knime.com/carstenlange/spaces/Workflows%20Related%20to%20the%20Texbook%20Practical%20Machine%20Learning%20with%20R%20by%20Carsten%20Lange/6.1%20Practical%20Machine%20Learning%20with%20R%20Tuning%20k-Nearest%20Neighbors%20(Wine%20Data)~YEZKZJbAvo-8U-6U/most-recent
- https://hub.knime.com/knime/spaces/Parameter%20Optimization%20Space/01_Classification/01_Parameter_Optimization_with_Nodes/03_Parameter_Optimization_Loop_on_Logit_with_CV~DpTZM1L0WrFHyb2t/current-state

Sampling strategies comparison

The initial analysis of the country-level data showed that the data from most countries are **imbalanced**, meaning that the data is not equally distributed between the two classes (religious and non-religious).

Class imbalance is a common issue that negatively impacts the classification performance of the model on the minority class because the model is overwhelmed by the majority class, and the results become biased. The imbalanced data combined with the small sample size causes a significant decline in model performance. In the context of classifying religiosity, it is important to have accurate
predictions, especially if the cost of misclassification is high, for example, if it

leads to negative consequences, lack of trust, or discrimination. We also need to monitor the overall performance of the model in predicting religiosity.

Example:

Table 6 presents the country-level data, with the proportion of minority to the majority class and the imbalanced learning metrics results on imbalanced data. The comparison of accuracy with the G-mean shows that while the accuracy is high in many countries, the models tend to perform significantly worse in the minority class. Hence the G-mean is low.

Table 6: Imbalanced data proportions and the classification results

Country	Religious(%)	Non-religious(%)	Accuracy	G-mean
China	13.44	86.56	0.88	0.49
India	88.38	11.62	0.88	0.18
Japan	23.42	76.58	0.74	0.23
Mexico	79.69	20.31	0.80	0.21
Nigeria	98.60	1.40	0.99	0.0

Table 5: Threshold metrics in supervised learning of imbalanced data

Metrics	Definition
Accuracy	The ratio of the correctly classified instances over the total number of classified instances
Precision	The proportion of instances that were labeled correctly among those with the positive label in the test data
Recall	The portion of positive instances in the test data that were labeled correctly
G-mean	The measure to maximize the accuracy of the model over each class by considering both classes for evaluation

When dealing with classification, it is necessary to address the challenges of imbalanced data and small sample size. Over-sampling methods are a group of resampling techniques used for imbalanced learning. The choice of resampling method depends on the data. Therefore, three methods were used to achieve the best performance on each data set.

1. A combination of random over-sampling and random under-sampling - generating new samples of the minority class by selecting the k-nearest neightboyrs of the minority sample and creating new samples on the line joining the minority sample and one of the neighbors. Next, apply an under-sampling technique to remove the majority samples near the minority sample and reduce the overlap and noise between the two classes

2. Synthetic minority over-sampling technique (SMOTE) is a widely used resampling technique that generates synthetic samples along the line segments that connect two samples of the minority class. (SMOTE is an effective technique because it generates new instances rather than duplicating the existing ones).
3. Adaptive Synthetic Sampling (ADA-SYN) is an extension of SMOTE which assigns a random value to the generated samples to have more variance in the synthetic data.

example:

> Table 7 presents the results of training the random forest on balanced data. It can be observed that the G-mean has improved substantially. The precision and recall indicate that the model is performing equally well in both classes, and high accuracy shows the overall effectiveness of the classification model.

Table 7: Classification results of balanced data

Country	Accuracy	Recall	Precision	G-mean
China	0.88	0.86	0.90	0.88
India	0.88	0.88	0.88	0.88
Japan	0.88	0.88	0.88	0.88
Mexico	0.88	0.88	0.88	0.88
Nigeria	0.98	0.97	0.99	0.98
Russia	0.82	0.81	0.81	0.82
Spain	0.84	0.81	0.87	0.81
Turkey	0.88	0.86	0.88	0.87
United States	0.83	0.85	0.81	0.84

Colunas a escolher:

B_COUNTRY
A_YEAR - year of survey - page 2
? Q_MODE - Mode of data collection
? G_TOWNSIZE
? H_SETTLEMENT - settlement type
? H_URBRURAL - urban or rural

	Country-specific list of codes in Annex
G_TOWNSIZE	**Settlement size_8 groups**
	Settlement size_8 groups 1.- Under 2,000 2.- 2,000-5,000 3.- 5,000-10,000 4.- 10,000-20,000 5.- 20,000-50,000 6.- 50,000-100,000 7.- 100,000-500,000 8.- 500,000 and more -5-.- No answer; Missing -4-.- Not asked in survey
G_TOWNSIZE2	**Settlement size_5 groups**
	Settlement size_5 groups 1.- Under 5,000 2.- 5000-20000 3.- 20000-100000 4.- 100000-500000 5.- 500000 and more -5-.- No answer; Missing -4-.- Not asked
H_SETTLEMENT	**Settlement type**
	Settlement type 1.- Capital city 2.- Regional center 3.- District center 4.- Another city, town (not a regional or district center) 5.- Village -4-.- Not asked -5-.- No answer; Missing
H_URBRURAL	**Urban-Rural**
	Urban-Rural Settlement type 1.- Urban 2.- Rural -5-.- No answer; Missing -4-.- Not asked

E1_LITERACY - Respondent's literacy

(wvs indexes)

Y001 - Post-materialist index 12-item

> Post-materialism is a term used in sociology and political science to describe a set of values and attitudes that prioritize non-materialistic or post-materialistic concerns over traditional materialistic concerns. This concept emerged in the late 20th century as societies in developed countries experienced economic prosperity and social changes that led to shifts in values and priorities.

Y002 - Post-Materialist index 4-item

Y003 - Autonomy Index

SACSECVAL => secular values - do not use

I_AUTHORITY - defiance - 1: Inverse respect for authority
I_NATIONALISM - defiance - 2: Inverse national pride

DISBELIEF - Inverse religious person - do not use

political views (kinda)
RELATIVISM
SCEPTICISM
AUTONOMY
EQUALITY
CHOICE
VOICE

RESEMAVAL - emancipative values

> "Emancipative value" refers to a specific type of cultural value or attitude that emphasizes the importance of individual autonomy, self-expression, freedom, and personal empowerment.

Q49 - satisfaction with your life
Q275 - education level
Q288 - income
Q262 - age
X003R - age recorded (6 intervals)

Y024 - VOICE
Y023 - CHOICE
Y022 - EQUALITY
Y021 - AUTONOMY
Y020 - RESEMAVAL
Y014 - SCEPTICISM
Y013 - RELATIVISM
Y012 - DISBELIEF
Y011 - DEFIANCE
Y001 - Post-Materialist index 12-item

X051 - Ethnic group
X047_WVS - Scale of incomes
X047R_WVS - Subjective income level (recoded in 3 groups)
X028 - Employment status
X025R - Education level (not present in one of the datasets)
X003 - Age
X003R - Age recoded (6 intervals)
X001 - Sex
S020 - Year survey
G006 - How proud of nationality

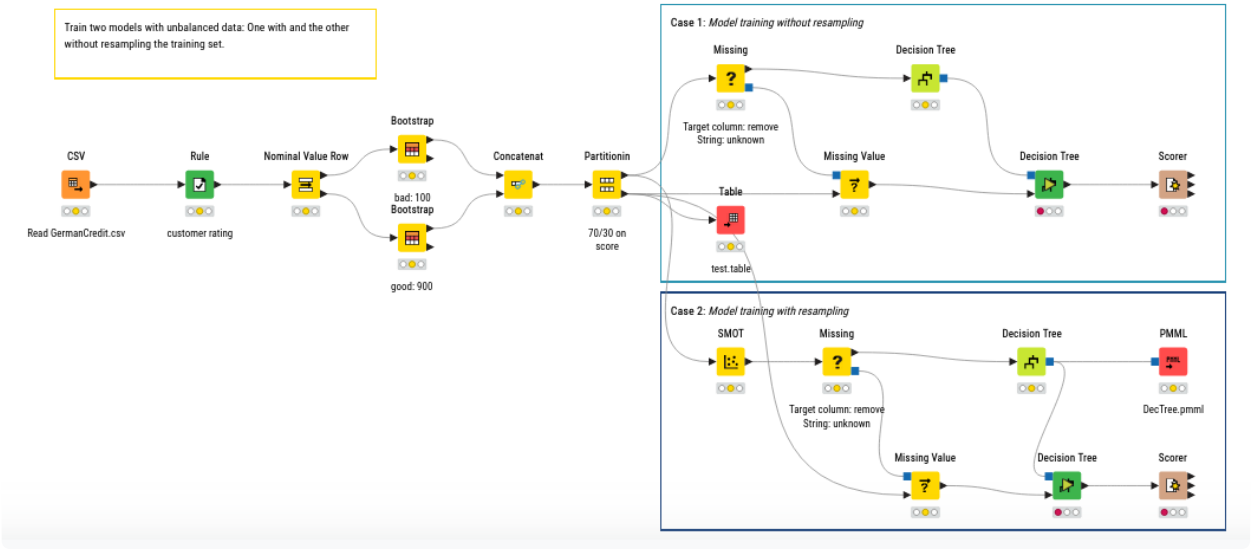
F115	Justifiable: Avoiding a fare on public transport	Q178	V199	V199	V205	V193	V297	V193
F116	Justifiable: Cheating on taxes	Q180	V201	V200	V206	V194	V298	V194
F117	Justifiable: Someone accepting a bribe	Q181	V202	V201	V207	V196	V306	V196
F118	Justifiable: Homosexuality	Q182	V203	V202	V208	V197	V307	V197
F119	Justifiable: Prostitution	Q183	V203A	V203	V209	V198	V308	V198
F120	Justifiable: Abortion	Q184	V204	V204	V210	V199	V309	V199
F121	Justifiable: Divorce	Q185	V205	V205	V211	V200	V310	V200
F122	Justifiable: Euthanasia	Q188	V207A	V206	V212	V201	V312	V201
F123	Justifiable: Suicide	Q187	V207	V207	V213	V202	V313	V202

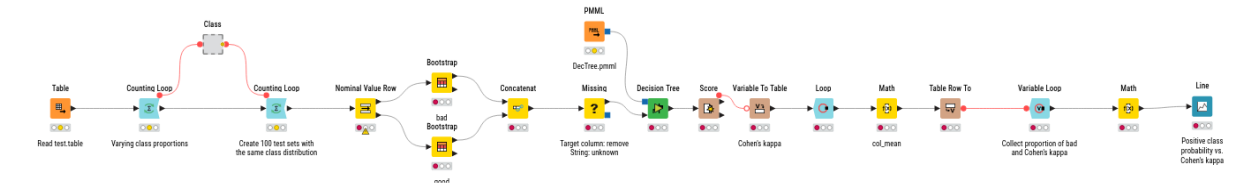
F050 - Believe in God
F034 - Religious Person
F028 - How often do you attend religious services
F025 - Religious denominations - major groups (like Q289)

E069_01	Confidence: Churches	Q64	V108	V131	V147	V135	V272	V135
E069_02	Confidence: Armed Forces	Q65	V109	V132	V148	V136	V273	V136
E069_03	Confidence: Education System	EVS_v117		V146_11			V274	
E069_04	Confidence: The Press	Q66	V110	V133	V149	V138	V276	V138
E069_05	Confidence: Labour Unions	Q68	V112	V135	V151	V140	V277	V140
E069_06	Confidence: The Police	Q69	V113	V136	V152	V141	V278	V141
E069_07	Confidence: Parliament	Q73	V117	V140	V155	V144	V279	V144
E069_08	Confidence: The Civil Services	Q74	V118	V141	V156	V145	V280	V145

COW_NUM - CoW country code numeric
A170 - Satisfaction with your life
A173 - How much freedom of choice and control
A008 - Feeling of happiness
A009 - State of health (subjective)

https://hub.knime.com/knime/spaces/Digital%20Healthcare/ECG%20Arrythmia%20Detection/ecg_cnn_mit~bWv0UtH6sTLAgcnn/most-recent





https://hub.knime.com/mlauber71/spaces/Public/_machine_learning_meta_collection~--zu7353LMafrYQ_/current-state

ROSE algorithm:
https://hub.knime.com/mlauber71/spaces/Public/kn_example_rose_balanced~xiWGmF_x4fuqKsK7/current-state

18 days later

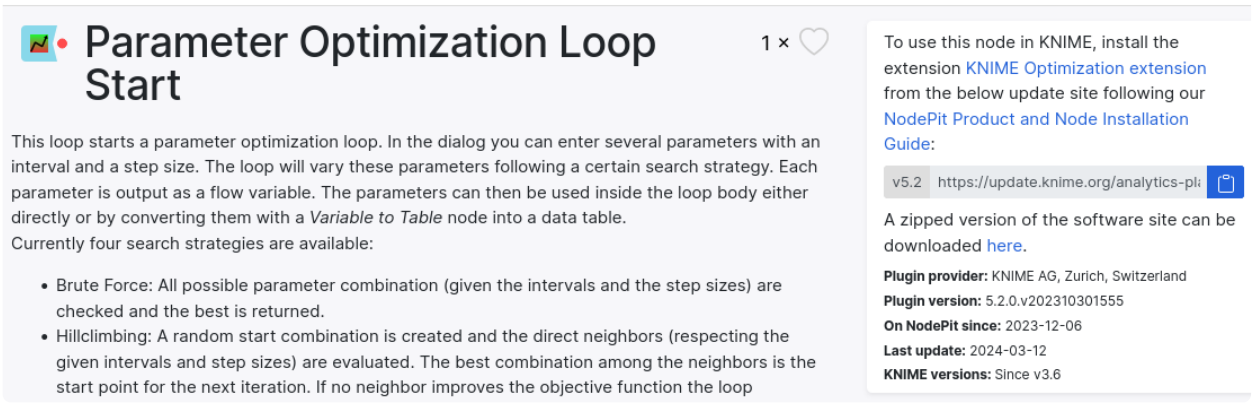
beginner Oct '17

Equal size sampling in my opinion isn't very useful as you loose a lot of data. SMOTE doesn't scale well if you have many features or simply doesn't work at all if you have ordinal features (integers which are either counts or "numerized categories")

In Knimes Random Forest / Tree ensemble nodes in Ensemble Configuration you can set your Data Sampling mode to Stratified. This ensures that each tree gets to see at least some rows of the minority class.

What KNIME lacks entirely is using class weights. What you could do hence is simply duplicate rows of the minority class (oversampling).Not ideal but might work (Duplicate with Concatenate Node as there is no oversampling node). But the duplication of course must happen on the training set only. Else you skew your Cross validation results.

Another option is to simply adjust the classification threshold or ditch the classification and work with the confidence output of the model (depends on use case, this works best when needing to prioritize work. Just do stuff with highest positive class value confidence first. eg. ranking by confidence for positive class value)



https://medium.com/low-code-for-advanced-data-science/99-accuracy-great-right-18eba5c7564d

note to self: ordenar primeiro por país e depois por ano e reduzir a average ou então colocar uma tabela de comparação depois