Università degli Studi di Milano

MSc. in Data Science and Economics

Statistical Learning Exam

A supervised and unsupervised analysis on the fertility rate among countries

Alessia Di Giovanni
alessia.digiovanni2@studenti.unimi.it

July 9, 2023

Abstract

This paper involves the analysis of a dataset that I built. A total of 143 countries are analysed, each characterised by 11 socio-economic and demographic variables. The analysis is divided into two parts. The first part conducts a supervised analysis with the fertility rate as the dependent variable. The first proposed method is a multiple linear regression and its diagnostics. The model is tested using k-fold cross-validation. Since there seems to be multicollinearity in the diagnostics, partial least square regression is proposed as an alternative solution. This is followed by regression using decision trees, both simple and random forest. Among all the supervised models, multiple linear regression shows better performance in terms of accuracy. The second part of the analysis focuses on unsupervised algorithms, particularly clustering. First, the k-means algorithm is performed solely on the numerical variables. Then, the grower distance is used to cluster the variables including the categorical variable of religion. In performing the clustering, the fertility rate is not used, not because it is the outcome, but rather to see if the naturally created clusters reflect different fertility rates. In the first method, 3 clusters are chosen, with one grouping developed countries and the other two comprising developing countries. In the second method, 2 clusters emerge, with one predominantly consisting of Islamic countries and the other predominantly consisting of Christian countries, including the more developed ones.

Contents

1 Problem statement

Before proceeding to search for a dataset, I wondered what topic would be interesting to investigate. Browsing through articles on the web, I came across some regarding the decline in fertility rates, particularly in developed countries. Total fertility rate of a population is the average of children that would be born to a woman over her lifetime if: she was to experience the exact current age-specific fertility rates (ASFRs) through her lifetime; she was to survive from childbirth until the end of her reproductive life. The ASFRs is of the relative frequency of childbearing among women of different ages within the reproductive years.

Once I had decided on my target variable, I started looking for variables that I thought might be conncected with the fertility rate. The 10 variables selected are as follows:

- **religion**. Religious beliefs can influence ideas about family, marriage and having children. Some religions may encourage larger families as a way of fulfilling religious duties or propagating the faith. Religious beliefs could also influence the division of labor within households. This can impact women's opportunities for education, employment and autonomy, which in turn can affect their reproductive choices;
- **life expectancy** (`age`): it is a statistical measure of the average time a human being is expected to live, based on the year of its birth, its current age, and other demographic factors including gender;
- **unemployment female rate** (`% of female labor force`). Unemployment refers to the share of the labor force that is without work but available for and seeking employment. Paradoxically, low unemployment rates can disguise substantial poverty in a country, while high unemployment rates can occur in countries with a high level of economic development and low rates of poverty. Countries with higher levels of development and employment opportunities for women, individuals may prioritize their careers and education before starting a family. This can lead to a delay in childbearing, resulting in lower fertility rates. As women participate more in the labor force and pursue higher education, they may choose to have children at a later stage in life, which can contribute to lower overall fertility rates. In the other hand unemployment often leads to economic instability and financial uncertainty. Couples who are unemployed or facing financial difficulties may delay or choose to have fewer children due to concerns about their ability to provide for their family's needs;
- **years of educational school** (`age`). The number of years of education can have an impact on the fertility rate. In countries where there is a higher emphasis on education and individuals tend to pursue advanced degrees or vocational training, women may prioritize their educational and career goals before starting a family. This focus on personal and professional development can contribute to a delay in entering parenthood until they have attained a certain level of economic stability and personal fulfillment. Moreover, higher levels of education often provide women with more opportunities and choices in life. With increased access to knowledge, healthcare and family planning resources, educated women may have a better understanding of the importance of family planning and the potential impact of having children on their personal and professional aspirations. This awareness, coupled with the desire to provide the best possible future for their children, may lead educated women to make deliberate decisions about the timing and number of children they have. Additionally, in more developed economies where education is highly valued, there may be greater availability of contraception, reproductive healthcare services, and supportive policies for work-life balance. These factors can further contribute to lower fertility rates as women have more control over their reproductive choices and can better align their family planning decisions with their educational and career trajectories;
- **use of contraceptives** (`% of women who are currently using, or whose sexual partner is currently using, at least one method of contraception, regardless of the method used`);
- **alcohol** (`total (recorded+unrecorded) alcohol per capita (15+) consumption`). Excessive alcohol consumption, particularly heavy and prolonged drinking, can have detrimental effects on both male and female fertility. In men, alcohol can disrupt sperm production, motility, and morphology, leading to reduced fertility. In women, alcohol can disrupt the hormonal balance, menstrual cycle regularity, and ovulation, making it more difficult to conceive;

- **number of days of paid maternity leave**(`days`);
- **freedom**(`from 0 to 100`). For each country and territory, Freedom in the World analyses the electoral process, political pluralism and participation, the functioning of the government, freedom of expression and of belief, associational and organizational rights, the rule of law, and personal autonomy and individual rights. Increased freedom can lead to various choices and opportunities for individuals, including the decision to have children. Freedom often goes hand in hand with access to healthcare, including reproductive healthcare and family planning services. When individuals have the freedom to make decisions about family planning and access to contraception, it can contribute to a decline in fertility rates.
- **migration rate** (`the difference between the number of persons entering and leaving a country during the year per 1,000 persons`). Migration can affect the age structure of a population. In some cases, migrants may be of reproductive age and have higher fertility rates compared to the host population. This can lead to a temporary increase in the overall fertility rate of the receiving country;
- **Gross National Income** (`thousand`). GNI stands for Gross National Income. It is a measure of the total income generated by a country's residents, including both domestic and foreign sources. GNI represents the total value of all goods and services produced within a country, including income earned from abroad (such as profits from foreign investments or remittances from citizens working in other countries), minus any income earned by foreign residents within the country. GNI is often used as an economic indicator to assess the overall economic performance and standard of living of a country.

I wanted to consider variables directly related to the fertility rate, such as contraceptive use and paid maternity days, variables related to a country's well being, such as years of education, alcohol consumption, life expectancy, variables related to a country's developing such as migration rate and finally two variables related to economy and culture, respectively GNI and the most widespread religion.

1.1 Dataset

For the construction of the dataset, the first part consists of scraping data from the following sources:

- https://www.kaggle.com/datasets/daniboy370/world-data-by-country-2020
- https://wisevoter.com/country-rankings/religion-by-country/
- https://data.worldbank.org/indicator/SL.UEM.TOTL.FE.NE.ZS
- https://www.worldeconomics.com/Indicator-Data/ESG/Social/Mean-Years-of-Schooling/
- https://www.un.org/development/desa/pd/data/world-contraceptive-use
- https://www.kaggle.com/datasets/sudhirnl7/human-development-index-hdi
- https://freedomhouse.org/countries/freedom-world/scores
- https://www.cia.gov/the-world-factbook/field/net-migration-rate/country-comparison

After merging these datasets, each one corresponding to a determinate variable, I obtain a dataset with 199 observations (different countries) and 11 features for each observation. There are 77 rows which contain missing values so I decide to drop ones with more than 2 missing values, 54 rows. In this way there are a few missing values, so I decide to fill them by using the mean of each predictor considered. My final dataset is given by 143 observations corresponding to 143 different countries and 12 variables (one label variable about countries' names and 11 features).

At this point the dataset is ready.

1.2 Pre-processing and EDA

First, the GNI is substituded by the log. Then I inspect the numeric variables. The table below report the min, max, mean and if there is the presence of outliers for each numeric variable.

	min	mean	max	presence of outliers
Fertility	1.10	2.25	5.90	NO
Life.expectancy	52.80	73.30	84.50	NO
Unemployment	0.50	6.45	35.90	YES
Alcohol	0.006	2.29	6.56	NO
Years.school	1.50	8.70	14.10	NO
Contraceptive	8.10	54.30	88.40	NO
Maternity.days	0.00	98.00	410.00	NO
Log_gni	6.37	9.25	11.77	NO
Freedom	3.00	56.00	100.0	NO
Migration.rate	-18.06	-0.63	13.01	YES

The presence of outliers can be noticed but since they are not incorrect observations, I decide not to omit them hoping they would not affect the analysis.
Looking at how fertility rate changes among religion, it can be seen that people who profess Islam are more likely to have a higher fertility rate, followed by those who profess Christianity, Buddhism and finally Hinduism. So the religion seems to be influent on fertility rate.

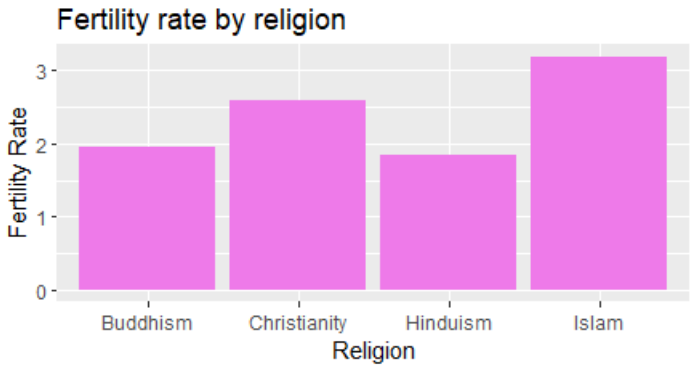


Figure 1: *Fertility rate by religion*

Before starting the analysis, I want to check the correlation between variables through the correlation matrix and in particular the correlation between my dependent variable and the independent ones.
As observed, the fertility rate shows a negative correlation with all other variables. The three variables that exhibit the strongest negative correlation are life expectancy, GNI and years of education. These variables are associated with the well-being of a country, confirming the hypothesis that in more developed or developing countries, the fertility rate tends to decrease. This is likely influenced by the use of contraceptives. It is surprising to see a negative correlation with maternity days, but it could be attributed to the fact that more days indicate more developed countries, which relates back to the previous discussion.

By examining the correlation matrix, we can also assess whether there is a problem of collinearity among the independent variables. It is notable that life expectancy, years of school, GNI and contraceptive use are strongly positively correlated with each other.

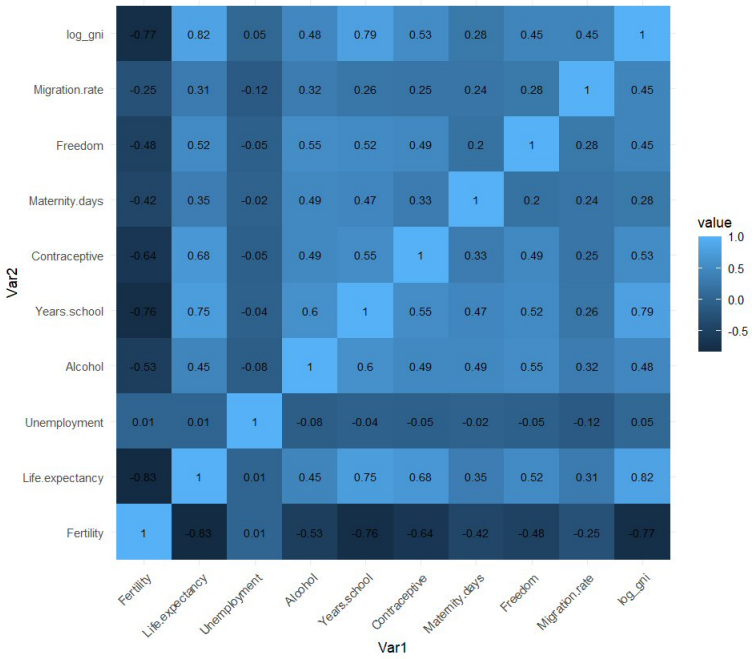


Figure 2: *Correlation matrix*

2 Supervised Learning[2]

2.1 Multiple linear regression

The first method I have chosen to investigate is multiple linear regression, which is a supervised technique. Linear regression is a valuable tool for predicting a quantitative response variable, such as the fertility rate. Since I have multiple predictors, I am performing a multiple linear regression analysis, which has the form

$$Y = \beta_0 + \beta_1 X_1 + +_2 X_2 + ... + \beta_p X_p + \epsilon \tag{1}$$

where X_j represents the jth predictor and β_j quantifies the association between that variable and the response. The β_js are interpreted as the average effect on Y of a one unit increase in X_j, holding all other predictors fixed. p is the number of the predictors, 11 in my analysis.

2.1.1 Target variable

First I focus on the variable `Fertility`. From the Shapiro-test and the QQ plot I can infer that it is not normally distributed. Trying using the log of it, normality is always refused but the situation seems improved, so I will use the log of Fertility as my dependent variable in the linear regression.

```
        Shapiro-Wilk normality test

data:  df$Fertility
W = 0.88623, p-value = 4.493e-09
```

Figure 3: *Shapiro-Wilk test for the variable* `Fertility`

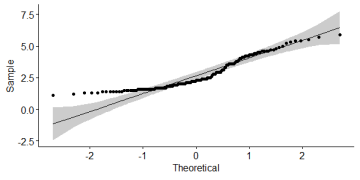


Figure 4: *QQ plot of the variable* `Fertility`

```
        Shapiro-Wilk normality test

data:  df$logFertility
W = 0.94369, p-value = 1.601e-05
```

Figure 5: *Shapiro-Wilk test for the variable* `logFertility`

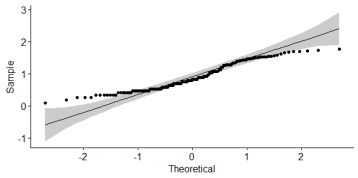
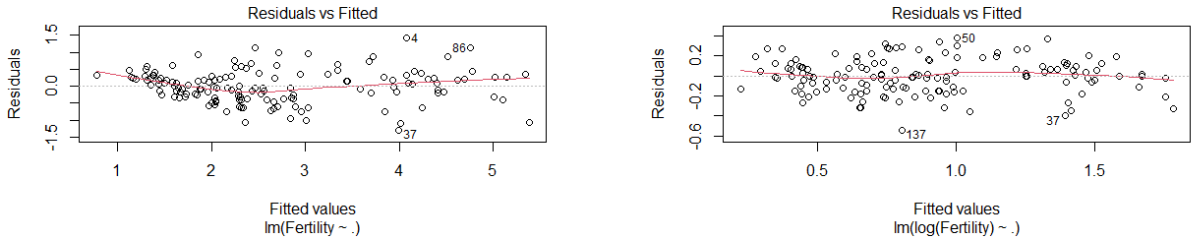


Figure 6: *QQ plot of the variable* `logFertility`

2.1.2 Diagnostic of the model

Linear relationship between the predictors and the response
The first assumption is that there exists a linear relationship between the predictors and the response variable. Residual plots can help us assess whether this assumption holds. These plots display the residuals, calculated as the difference between the observed response values and the predicted values, $e_i = y_i - \hat{y}_i$, against the predictor variables, x_i. In an ideal case, the red line representing the residuals should be horizontal. By examining the residual plots for two linear regressions — one using the original fertility rate as the dependent variable and the other using the logarithm of the fertility rate — we can make comparisons. The residual plot for the regression with the logarithm of fertility rate as the dependent variable shows a more horizontal red line compared to the plot using the original fertility rate. This suggests that the second regression model provides a better fit to the data compared to the first one.



(a) (b)

Figure 7: Residuals vs Fitted values

From this point onwards, only the regression using the logarithm of fertility as the dependent variable is considered.

Homoscedasticity/Heteroscedasticity
A second assumption is that the error terms have a constant variance, $Var(\epsilon_i) = \sigma^2$, so the presence of homoscedasticity. Homoskedasticity occurs when the variance of the errors is consistent for all observations; in contrast heteroskedasticity refers to a situation where the variance of the errors in a regression model is not constant across all observations. In the presence of heteroskedasticity, there are two main implications for the least squares estimators:

- the least squares estimator remains a linear and unbiased estimator, but it is no longer the most efficient estimator. In other words, there exists another estimator that has a smaller variance, meaning it provides more precise estimates;
- the standard errors computed for the least squares estimators are incorrect. This has consequences for confidence intervals and hypothesis testing that rely on these standard errors. Incorrect standard errors can lead to misleading conclusions, affecting the interpretation of statistical significance and the validity of statistical inference.

Therefore, addressing heteroskedasticity is crucial to ensure reliable and accurate statistical analysis, as it allows for obtaining more efficient estimators and obtaining valid statistical inferences from the regression model. One way to

visually detect whether heteroscedasticity is present is to observe the plot of the residuals against the fitted values of the regression model. If the residuals become more spread out at higher values in the plot, this is a tell-tale sign that heteroscedasticity is present. This not seems my case. To get a confirm, I run the Non-costant Variance Score Test and the Bresuch-Pagan test: since the p-values are above 0.05 for both tests, the homoscedasticity hypothesis is accepted.

```
        studentized Breusch-Pagan test

data:  reg1
BP = 19.016, df = 12, p-value = 0.08813
```

(a)

```
Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 1.409889, Df = 1, p = 0.23507
```

(b)

Figure 8: Tests for homoscedasticity

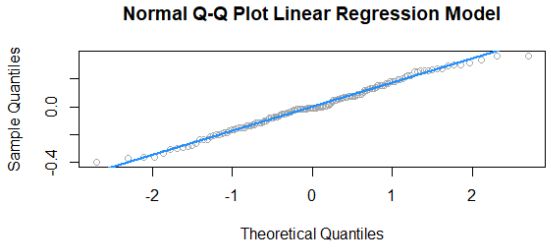
Normality of the residuals

The third assumption is that residual errors are assumed to be normally distributed. To test for it, the Shapiro-Wilk test is run. Since the p-value is above 0.05, the null hypothesis is accepted: the residuals are sampled from a normal distribution. This could be confirmed by looking at the corresponding Q-Q plot.

```
        Shapiro-Wilk normality test

data:  resid(reg1)
W = 0.99017, p-value = 0.4163
```

(a)



(b)

Figure 9: Normality of the residuals

Multicollinearity

With multicollinearity, the regression coefficients are still consistent but are no longer reliable since the standard errors are inflated. It means that the model's predictive power is not reduced, but the coefficients may not be statistically significant with a Type II error. With the correlation matrix, I can see which variables are correlated each other. To test the multicollinearity I inspect the variance inflation factors. A tolerance < 0.1 might indicate multicollinearity. A VIF exceeding 5 requires further investigation, whereas VIFs above 10 indicate multicollinearity. Ideally, the Variance Inflation Factors are below 3. In the table it can be seen that some VIFs are bigger than 3 so maybe there would be multicollinearity. VIFs between 1 and 5 there is moderate correlation between a given predictor variable and other predictor variables in the model.

	Variables	Tolerance	VIF
1	ReligionChristianity	0.2123626	4.708926
2	ReligionHinduism	0.7259949	1.377420
3	ReligionIslam	0.2143834	4.664540
4	Life.expectancy	0.2317055	4.315823
5	Unemployment	0.8899327	1.123681
6	Alcohol	0.3469966	2.881873
7	Years.school	0.2192209	4.561609
8	Contraceptive	0.4362631	2.292195
9	Maternity.days	0.7503384	1.332732
10	Freedom	0.5121839	1.952424
11	Migration.rate	0.7654961	1.306342
12	log_gni	0.2050809	4.876125

Figure 10: *Variance Inflation Indicators*

Outliers and leverage points

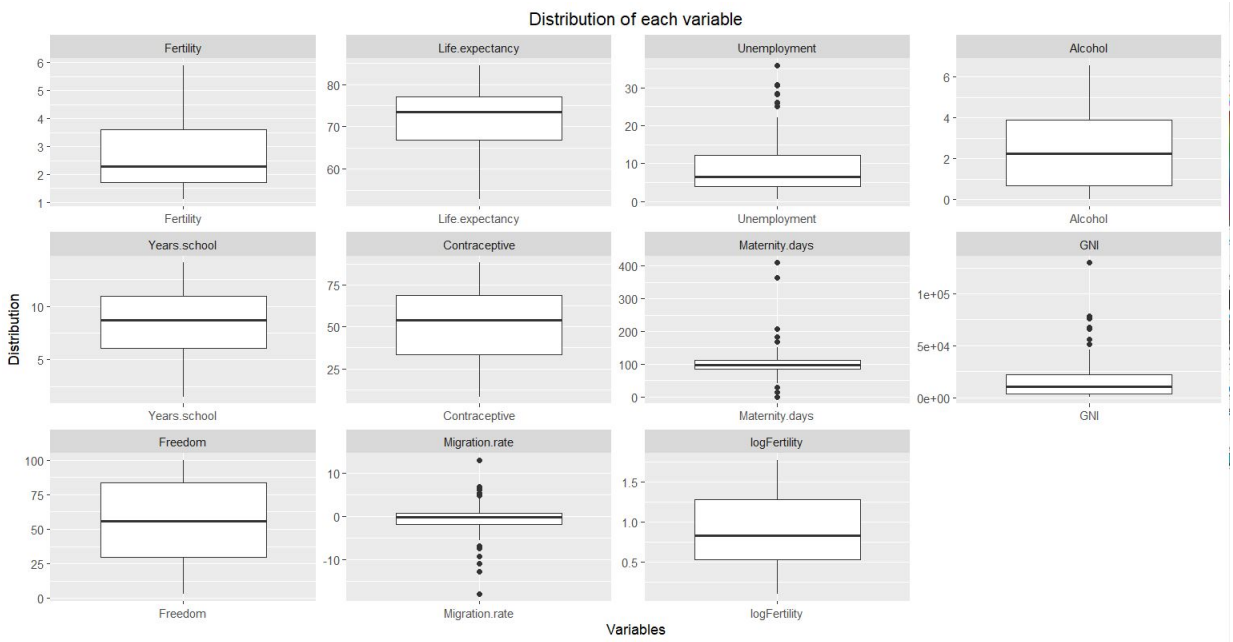


Figure 11: *Boxplots of each variable*

An outlier is a point for which y_i is far from the value predicted by the model. Observations with high leverage have an unusual value for x_i. The residuals vs leverage plot is a useful tool for identifying outliers and high leverage points in a regression analysis.

- Outliers appear as points that have large vertical distances from the horizontal reference line at zero. These points indicate data points with extreme y-values that deviate significantly from the overall trend;
- high leverage points are identified by their position on the horizontal axis. Points that are located towards the edges of the plot have high leverage because they have extreme x-values that can strongly influence the regression line.

From the plot, it can be seen that there are both outliers and leverage points.

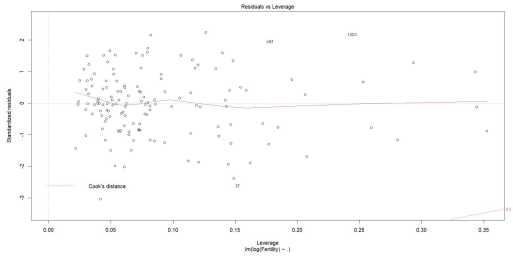


Figure 12: *Residuals vs leverage*

But plotting the studentized residuals, computed by dividing each residual e_i by its estimated standard error, it can be seen that the values of outliers are somehow acceptable (the absolute value of studentized residuals is less than 3).

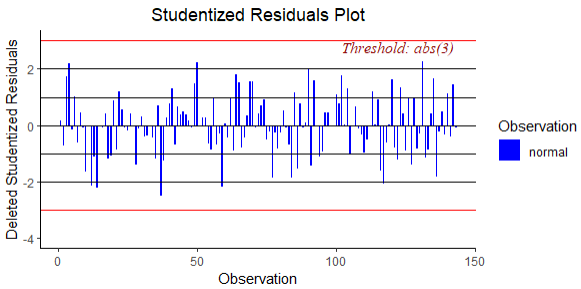


Figure 13: *Studentized residuals*

Leverage points can be detected by examining the leverage statistic or the hat-value. A value of this statistic above $2(\frac{p+1}{n})$ indicates an observation with high leverage, where, p is the number of predictors and n is the number of observations.

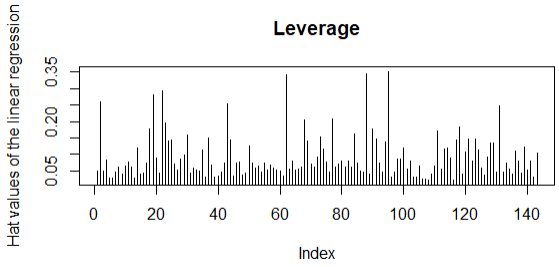


Figure 14: *Leverage*

It can be seen that there are a lot of leverage points, but I can't consider them influential value. An influential value is a value, which inclusion or exclusion can alter the results of the regression analysis. In residuals and leverage plot, the data don't present any influential points. Cook's distance lines (red dashed lines) are shown on the residuals vs leverage plot but all points are well inside of the Cook's distance lines.

2.1.3 Results

Coefficient estimates on the whole dataset are displayed in Figure 15. The regressors which are significant at 99% confidence are: `ReligionChristianity`, `ReligionHinduism`, `ReligionIslam`, `Life.expectancy`, `Contraceptive`, `Maternity days`, `logGni` and `Migration.rate`. TChristianity has a positive impact on the fertility rate as it places significant importance on the value of building a family. Similarly, the migration rate has a positive influence. On the other hand, all other variables seem to have a negative influence. This was expected for variables related to the well-being of the country, while the negative impact of days of maternity leave is counter-intuitive. The overall quality of the model can be assessed by examining the R-squared (R^2) and Residual Standard Error (RSE). In multiple linear regression, the R^2 or Coefficient of Determination measures the proportion of the variance in the dependent variable that can be explained by the independent variables in the model. It ranges from 0 to 1, where a value of 1 indicates that the model explains all the variability in the data, and a value of 0 indicates that the model does not explain any of the variability. A higher R-squared value indicates a better fit of the model to the data. The RSE estimate gives a measure of error of prediction. The lower the RSE, the more accurate the model.

Adjusted R^2	RSE
0.83	0.07

```
Call:
lm(formula = log(Fertility) ~ ., data = df_reg)

Residuals:
     Min       1Q   Median       3Q      Max
-0.39855 -0.11538 -0.00698  0.11774  0.36871

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)           3.9517253  0.2424066  16.302  < 2e-16 ***
ReligionChristianity  0.1560815  0.0671676   2.324 0.021690 *
ReligionHinduism     -0.2422432  0.1214190  -1.995 0.048123 *
ReligionIslam         0.2269408  0.0719008   3.156 0.001986 **
Life.expectancy      -0.0237649  0.0041580  -5.716 7.12e-08 ***
Unemployment         -0.0007014  0.0021849  -0.321 0.748697
Alcohol              -0.0020374  0.0133229  -0.153 0.878697
Years.school         -0.0099604  0.0100631  -0.990 0.324109
Contraceptive        -0.0022256  0.0010597  -2.100 0.037647 *
Maternity.days       -0.0012528  0.0003297  -3.800 0.000222 ***
Freedom               0.0002488  0.0007120   0.349 0.727337
Migration.rate        0.0126548  0.0046986   2.693 0.008007 **
log_gni              -0.1280434  0.0271846  -4.710 6.26e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1773 on 130 degrees of freedom
Multiple R-squared:  0.8439,    Adjusted R-squared:  0.8295
F-statistic: 58.59 on 12 and 130 DF,  p-value: < 2.2e-16
```

Figure 15: *Results*

2.2 Resampling method: k-fold cross validation

Resampling methods are techniques used in statistical analysis to estimate the performance of a model or to validate its predictive ability. Instead of relying solely on a single training set and test set, resampling methods involve repeatedly drawing samples from the available data to obtain multiple estimates of model performance. An example of ensamble method is cross-validation, used to estimate the test error associated with a given statistical learning method. Through cross validation the performance of the model is assessed. The algorithm of k-fold cross validation is:

- randomly divide the data set into k subsets (in general k=5 or k=10);
- reserve one subset as the test set and train the model on the remaining k-1 subsets;
- evaluate the model's performance by testing it on the reserved subset and record the mean squared error;
- repeat steps 2 and 3, each time using a different subset as the test set;
- calculate the average of the mean squared errors, which serves as the cross-validation error and represents the model's performance metric.

Choosing the appropriate value of k is important. A lower value of k introduces more bias, which is undesirable; on the other hand, a higher value of k reduces bias but may result in larger variability. It is worth noting that a smaller value of k, such as k = 2, resembles the validation set approach, while a higher value of k, such as k = n, resembles the LOOCV approach, where validation set and LOOCV are two other cross validation approaches. I decide to use k-fold cross validation to test the performance of the multiple linear regression. The results are shown below.

```
Linear Regression

143 samples
 10 predictor

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 128, 130, 129, 127, 129, 127, ...
Resampling results:

  RMSE       Rsquared  MAE
  0.1874469  0.835699  0.1498598

Tuning parameter 'intercept' was held constant at a value of TRUE
```

Figure 16: *K-fold cross validation for multiple linear regression*

The metrics used to evalute the performance are: RMSE (Root Mean Squared Error), R^2 and MAE (Mean Absolute Error).

- RMSE measures the average magnitude of the residuals, which are the differences between the predicted values and the actual values. It provides an indication of how well the model's predictions align with the true values. A lower RMSE indicates better model performance, as it represents a smaller average prediction error. RMSE = 0.19 means a very good model performance;

- $R^2 = 0.84$ indicates a good fit of the model to the data;
- MAE measures the average absolute difference between the predicted values and the actual values. It provides a similar measure of prediction accuracy as RMSE but without considering the squared differences. A lower MAE indicates better model performance, as it represents a smaller average absolute prediction error. Also in this case I can be satisfied from a value of MAE equal to 0.14.

The reported values represent the average performance over the k iterations of the cross-validation process.

R^2	RMSE
0.84	0.19

2.3 Dimension reduction methods: partial least square regression

As it is seen above, there could be some problems due to the multicollinearity. Dimension reduction methods could be useful in this case, as they work summarizing the original predictors into few new variables called principal components (PCs), which are then used as predictors to fit the linear regression model. Two different approaches are the principal components regression and the partial least square regression. But while PCR involves identifying linear combinations, or directions, that best represents the predictors in an unsupervised way, so not considering the response variable, the PLS is a supervised dimension reduction method which makes use of the response variable in order to identify new features that not only approximate the old features well, but also that are related to the response. For this reason, in order to solve the problem of multicollinearity, I decide to run a partial least square regression. The dataset is splitting in training set (80% of the dataset) and test set (the remaining 20%). The partial least squares regression is run on the training set to find a set of latent variables (components) that capture the maximum covariance between the predictors and the response variable. Variables are standardized, making them comparable and reducing the impact of differences in scale. Cross-validation is used to identify the optimal number of principal components to be incorporated in the model, finding out when the cross-validation RMSE is minimized. The optimal number of principal components included in the PLS model is 2. This captures 56% of the variation in the predictors and 82% of the variation in the outcome variable.

```
Data:   X dimension: 116 9
        Y dimension: 116 1
Fit method: kernelpls
Number of components considered: 9

VALIDATION: RMSEP
Cross-validated using 10 random segments.
       (Intercept)  1 comps  2 comps  3 comps  4 comps  5 comps  6 comps  7 comps  8 comps  9 comps
CV           1.238   0.6427   0.5403   0.5485   0.5461   0.5516   0.5524   0.5526   0.5527   0.5527
adjCV        1.238   0.6417   0.5384   0.5459   0.5438   0.5489   0.5496   0.5498   0.5499   0.5499

TRAINING: % variance explained
           1 comps  2 comps  3 comps  4 comps  5 comps  6 comps  7 comps  8 comps  9 comps
X            44.86    56.00    64.56    72.74    78.62    81.95    91.69    96.91   100.00
Fertility    74.12    82.87    83.18    83.24    83.26    83.27    83.27    83.27    83.27
```

Figure 17: *PLS results*

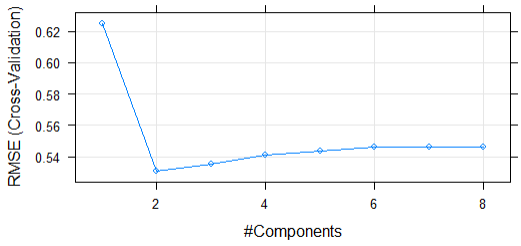


Figure 18: *Plot of number of components versus RMSE*

Making predictions on the test set, the results obtained for RMSE and R^2 are as follow:

R^2	RMSE
0.72	0.64

It can be seen that the R^2 in PLS is less than the one in the multiple linear regression as the RMSE is bigger, so the performance of multiple linear regression is better.

2.4 Tree-based methods

Three-base methods are supervised techniques used for regression. Their advantage lies in being easy to interpret. However, they are not as competitive in terms of prediction accuracy when compared to other supervised techniques. For this reason, I will also consider random forests, which combine a large number of trees, improving prediction accuracy at the expense of interpretability.

2.4.1 Regression trees

Tree-based algorithms for regression involve partitioning the predictor space into distinct and non-overlapping regions. Each observation falling within a region is assigned the same prediction, which is the mean of the training observations within that region. The predictor space is divided into high-dimensional rectangles (boxes) $R_1, .., R_j$ that minimize the residual sum of squares (RSS), given by

$$\sum_{j=1}^{J} \sum_{i \in R_j} (y_i - \hat{y_{R_j}})^2. \tag{2}$$

To construct the tree, a top-down, greedy approach known as recursive binary splitting is employed. It starts at the root node and iteratively splits the predictor space into two new branches based on the selected splitting criterion. This process continues until a stopping criterion is met.
First I build two trees on the entire dataset, using the `ctree` and the `rpart` function. If we look at the tree built using `rpart`, the most predictive variable is life expectancy with a threshold of 69 years: life expectancy is in fact the most correlated variable with fertility with a coefficient of correlation equal to -0.82. The other variables involved are years of school, log of Gross National Income and maternity days, all variables relative to the well-being of the countries. The values at the leaf nodes in the decision tree represent the average fertility rate. The analysis shows that as life expectancy increases, the fertility rate tends to decrease. This relationship can be attributed to various factors such as socioeconomic development. Countries with lower life expectancy often have less developed infrastructure and limited opportunities for women to pursue careers. Additionally, healthcare systems in these countries may be less advanced, resulting in higher fertility rates and fewer opportunities for family planning. The variables such as year of schooling, GNI and maternity days also play a role in determining the fertility rate. Higher values for these variables are associated with lower fertility rates. This suggests that factors such as education, economic prosperity and supportive maternity policies can contribute to lower fertility rates.
To evaluate the accuracy of the tree-based method, the decision tree is rebuilt using only the training set and predictions are made using the test set. The mean squared error obtained is 0.49 and the root mean squared error is 0.70.

RMSE
0.70

These error values are higher compared to the other two supervised methods, indicating that the tree-based method may have slightly lower predictive accuracy in this case.

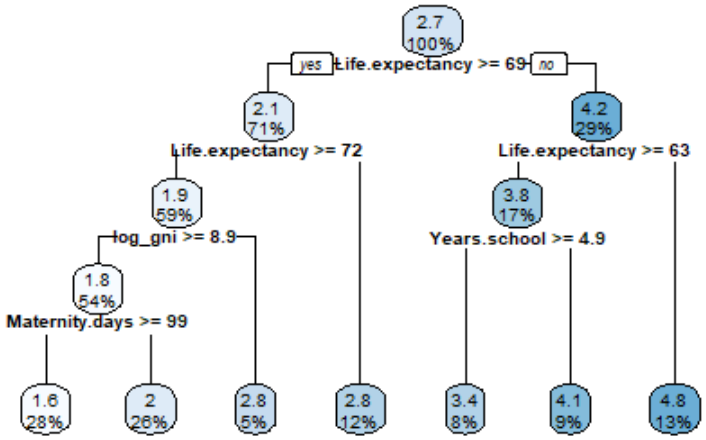


Figure 19: *Tree built using rpart function on the entire dataset*

2.4.2 Tree pruning

If the tree is too complex, the predictions on the training set could be good, but there may be overfitting and so there could be poor test set performance. Taking a smaller tree, we can have lower variance and better interpretation, at the cost of little bias. To do this the tree can be pruned. To decide where to cut the tree, the complexity parameter based on the lowest cross-validation error (x-error) from the cp table is selected. The best cp corresponds to five splits, that is the same number used in the original tree. I see that taking cp = 0.022120 has a lower x-error, so I can try to use only 3 splits and seeing if the situation is improved. Calculating the RMSE, it results equal to 0.69, meaning that in this case pruning the tree improve a little bit the performance.

RMSE
0.69

	CP	nsplit	rel error	xerror	xstd
1	0.696072	0	1.00000	1.02646	0.116534
2	0.076148	1	0.30393	0.41319	0.056597
3	0.028989	2	0.22778	0.37091	0.052999
4	0.022120	3	0.19879	0.34915	0.049202
5	0.013687	4	0.17667	0.33810	0.049100
6	0.010000	5	0.16298	0.33009	0.049426

Figure 20: *CP table*

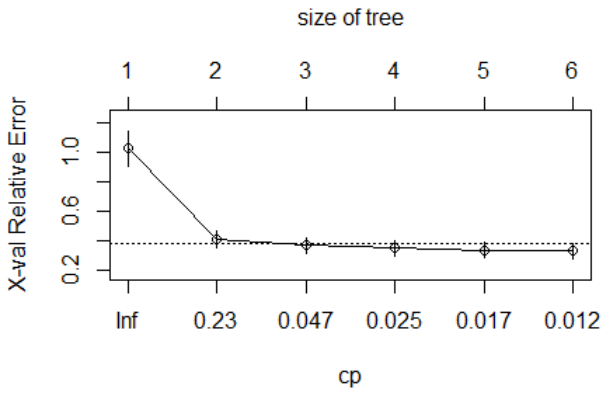


Figure 21: *CP plot*

2.4.3 Bagging and random forest

Using a single decision tree can lead to the problem of high variance. To reduce the variance, bagging (bootstrap aggregating) can be used. The algorithm works as follows:

- B bootrapped samples are taken from the dataset;
- for each sample a decision tree is built;
- the average of the predictions for each tree is considered for the final model.

Averaging a set of observations reduce variance. In our case we construct B regression trees using B bootstrapped training sets and then average the resulting predictions. Each individual tree has high variance but low bias; averaging these B trees reduces the variance. Bagging a large number of trees, the result statistical learning procedure can't be represented: bagging improves prediction at the expense of interpretability. All the 11 predictors are used in bagging and the final RMSE results 0.64.

RMSE
0.64

To improve bagged trees, random forest can be used. As in bagging, a number of decision trees on bootstrapped training samples are built. The difference with bagging is that for each split only a random selection of m out of p predictors is considered (in general m is equal to the square root of the total number of predictors). In this way the correlation among the trees is reduced, leading potentially better performance. Running the random forest model, the RMSE obtains is equal to 0.55.

RMSE
0.55

The variance importance plot indicates that across all of the tree considered in the random forest, life expectancy and GNI are the two most important variables, followed by the years of school, use of contraceptive and consumption of alcohol.

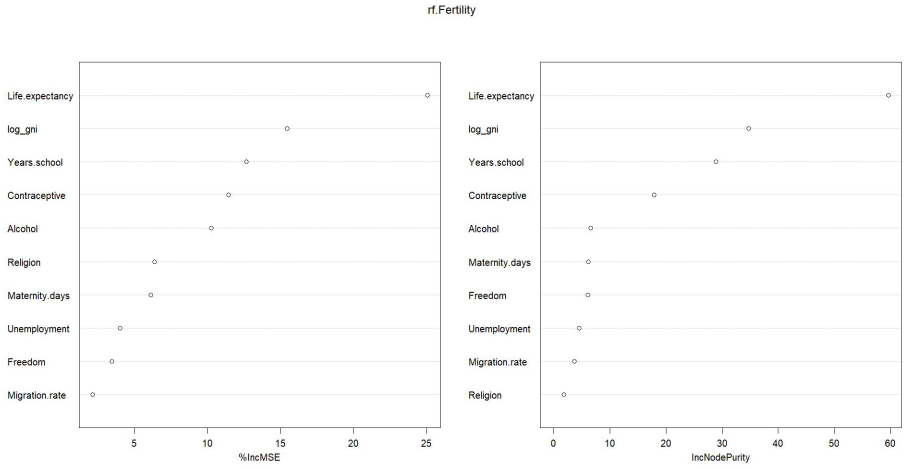


Figure 22: *Importance of variables in random forest*

3 Unsupervised learning

3.1 Clustering methods

Clustering refers to a set of unsupervised techniques for finding subgroups in a dataset. The goal is to find clusters so that observations within each group are similar to each other, while observations in different groups are different.

3.1.1 Numerical variables - K-means algorithm

For this first part of clustering analysis, I have decided to consider only the numerical variables and exclude the categorical one, the `Religion`. When the variables are measured in different units or have different scales, it is recommended to scale the dataset to avoid distortions caused by these differences. Therefore, before applying the clustering algorithm, I will standardize the data. Standardization transforms the variables so that they have a mean of 0 and a standard deviation of 1. The variable `Fertility` is not used: this will not be the outcome, because this is an unsupervised analysis, but I want to see if the clusters that will be naturally computed, will have also a different rate of fertility, meaning that the variables chosen are a good indices on how the fertility rates changes among the countries.
Using the k-means algorithm, the dataset is partitioned into k distinct and non-overlapping clusters, where k is a predetermined number. K points are selected as cluster centroids and each data point is assigned to the cluster centroid that is closest to it based on a distance metric, typically the Euclidean distance (this is why k-means algorithm can be applied only to numerical variables). The centroids are then recalculated by taking the average of all data points within each cluster. This process is repeated until convergence is achieved. The goal is to minimize the within-cluster variance, ensuring that the data points within each cluster are similar to each other.
First, I have to choose the number k. The `wssplot` function generates a plot that displays the within-cluster sum of squares (WCSS) for different numbers of clusters in a k-means clustering analysis. WCSS represents the sum of squared distances between each data point and its assigned centroid within each cluster. The plot helps in determining the appropriate number of clusters to use in the k-means algorithm. By examining the plot, I can identify the number of clusters where the decrease in WCSS slows down significantly, indicating a good balance between cluster compactness and the number of clusters. To have confirmation about the number of clusters chosen, I decide to use also the `clusGap` function to compute the gap statistic for clustering. It estimates the optimal number of clusters by comparing the within-cluster dispersion of the observed data to that of reference null datasets. From both plots it seems convenient to choose 3 as the optimal number of clusters.

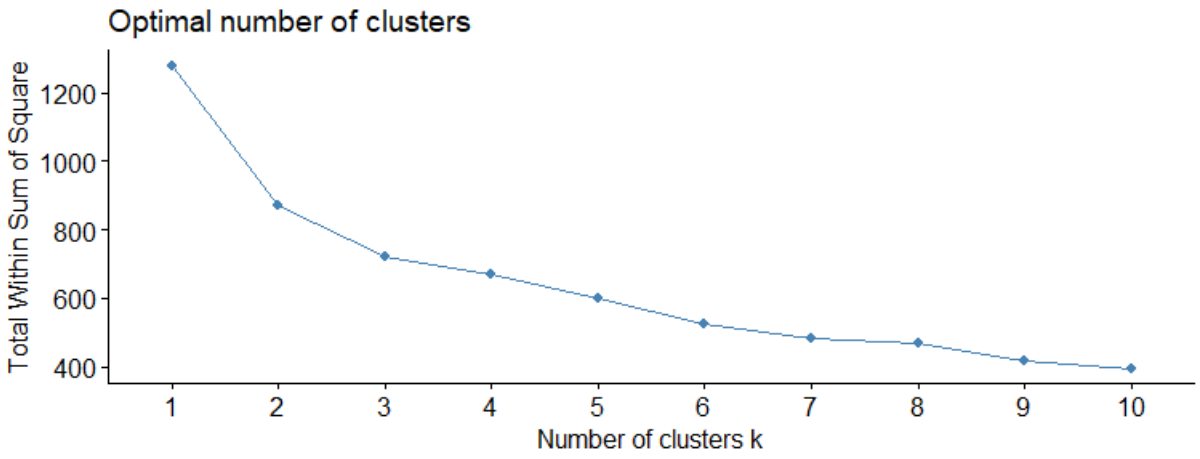


Figure 23: *Wss plot*

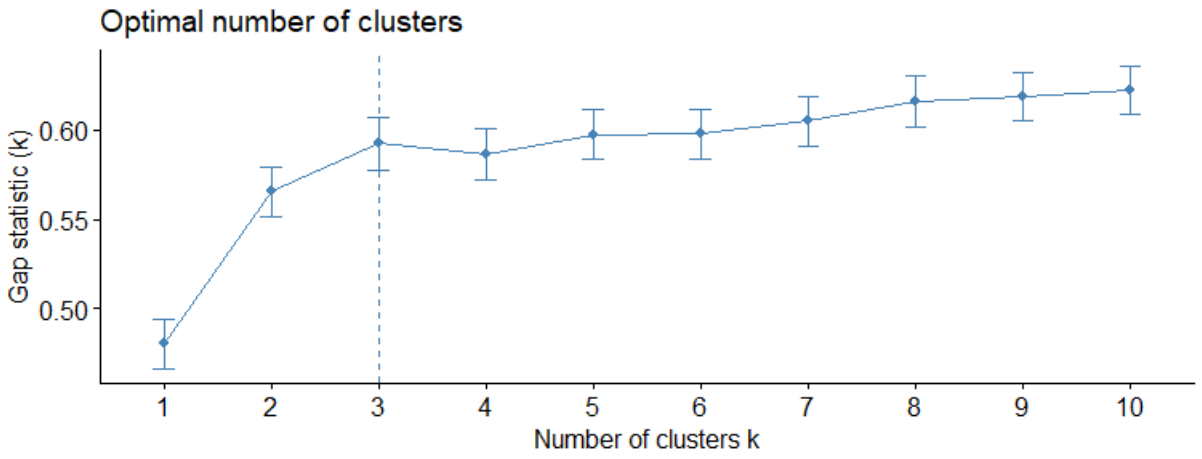


Figure 24: *Gap statistic*

The clusters obtained presented these means for all the numerical variables.

	Cluster 1	Cluster 2	Cluster 3
Fertility	1.65	2.35	4.20
Life.expectancy	79.03	73.81	63.18
Unemployment	7.18	10.73	8.71
Alcohol	4.49	1.61	1.40
Years.school	11.53	8.68	4.75
Contraceptive	66.61	54.88	31.45
Maternity.days	135.97	82.64	84.93
Log_gni	10.20	9.38	7.66
Freedom	82.88	47.58	38.55
Migration.rate	1.31	-2.03	-0.81

N.B.:the fertility rate did not influence the creation of the clusters but was calculated once the dataset was divided into the three clusters. It can be seen that the fertility rates changes among the three clusters, meaning that the natural clusters obtained are good. Let's see which countries belong to each cluster.

Cluster 1	Albania, Antigua and Barbuda, Argentina, Australia, Austria, Barbados, Belarus, Belgium, Bosnia and Herzegovina, Brazil, Bulgaria, Canada, Croatia, Denmark, Estonia, Finland, France, Georgia, Germany, Greece, Hungary, Ireland, Italy, Japan, Latvia, Lithuania, Malta, Mongolia, Netherlands, New Zealand, Norway, Panama, Poland, Portugal, Romania, Serbia, Singapore, Slovenia, Spain, Sweden, Switzerland, Trinidad and Tobago, United Kingdom, Uruguay
Cluster 2	Algeria, Armenia, Azerbaijan, Bahrain, Bangladesh, Belize, Bhutan, Botswana, China, Colombia, Costa Rica, Cuba, Dominica, Dominican Republic, Ecuador, Egypt, El Salvador, Fiji, Gabon, Grenada, Guatemala, Guyana, Honduras, India, Indonesia, Iraq, Jamaica, Jordan, Kazakhstan, Kuwait, Kyrgyzstan, Lebanon, Libya, Malaysia, Maldives, Mauritius, Mexico, Montenegro, Morocco, Nicaragua, Oman, Paraguay, Peru, Philippines, Qatar, Saudi Arabia, South Africa, Sri Lanka, Suriname, Thailand, Tonga, Tunisia, Turkey, United Arab Emirates, Uzbekistan, Vanuatu
Cluster 3	Afghanistan, Angola, Benin, Burkina Faso, Burundi, Cambodia, Cameroon, Central African Republic, Chad, Comoros, Djibouti, Equatorial Guinea, Eritrea, Ethiopia, Ghana, Guinea, Haiti, Kenya, Kiribati, Lesotho, Liberia, Madagascar, Malawi, Mali, Mauritania, Mozambique, Myanmar, Nepal, Niger, Nigeria, Pakistan, Papua New Guinea, Rwanda, Senegal, Sierra Leone, Solomon Islands, Sudan, Tajikistan, Togo, Uganda, Yemen, Zambia, Zimbabwe

As can be observed, countries belonging to Cluster 1 are mostly European and American. They represent the most developed countries, as can be seen from variables related to the well-being of the country: life expectancy is high, as well as the number of years dedicated to education and the use of contraceptives. In these countries, the level of freedom is also very high. On the other hand, Clusters 2 and 3 consist of a large number of African countries and are predominantly developing countries. In Cluster 3, in particular, the fertility rate is very high, while the level of freedom is very low.

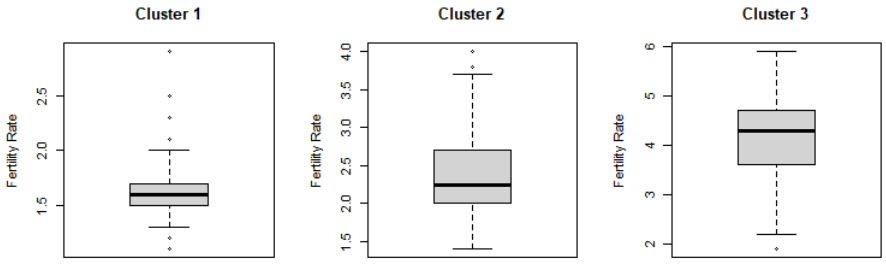


Figure 25: *Boxplot for fertility rates*

A way to evaluate the performance of the k-means clustering method is by a silhouette analysis. The silhouette coefficient ranges from -1 to 1, where a value close to 1 indicates that the data points are well-clustered and separated, a value close to 0 indicates overlapping clusters and a value close to -1 indicates that the data points may have been assigned to the wrong clusters. Based on my results, it can be said that clustering achieved reasonable cluster separation, as the average silhouette is 0.23.

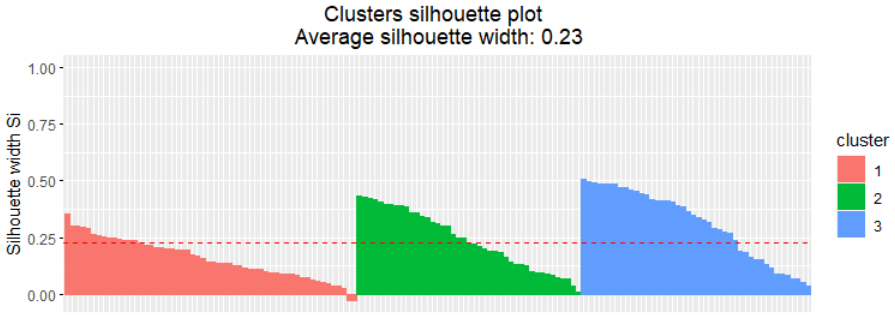


Figure 26: *Silhouette analysis*

3.1.2 Numerical and categorical variables - the gower distance[1]

Last analysis I want to explore is a clustering method using not only the numerical variables but considering also the categorical variable `Religion`. To implement clustering on both numerical and categorical variables, the gower distance can be used. The gower distance is calculated as the average of partial dissimilarities between instances. The gower index formula takes into account the feasibility of comparing two observations and calculates the dissimilarity accordingly. For qualitative variables, the gower index is based on whether the observations fall into the same class. For quantitative variables, the gower index considers the difference between the values of the variables. The gower distance is bounded between 0 and 1. A distance of 1 indicates perfect similarity, where all observations fit into the same classes and have equal quantitative values. A distance of 0 indicates no pair of observations fitting into the same classes and having similar values.
A matrix of gower distances is calculated. There are 10153 dissimilarities with a minimum of 0.02294 and a maximum of 0.68354. Below the most similar pair and the most dissimilar one.

```
  Country     Religion Fertility Life.expectancy Unemployment Alcohol Years.school
Lithuania Christianity       1.6           75.7          6.6    6.56         13.0
   Latvia Christianity       1.6           75.2          6.5    6.11         12.8
Contraceptive Maternity.days Freedom Migration.rate  log_gni
         68.8            126      89          -4.34 10.16608
         67.8            112      88          -5.06 10.02522
```

Figure 27: *Most similar pair of observations*

```
Country     Religion Fertility Life.expectancy Unemployment Alcohol Years.school Contraceptive
Finland Christianity       1.4           81.7          7.0    5.05         12.4          85.5
   Chad        Islam       5.7           54.0          0.8    0.50          2.3           8.1
Maternity.days Freedom Migration.rate   log_gni
           147     100           2.25 10.567927
            98      15          -0.12  7.596392
```

Figure 28: *Most dissimilar pair of observations*

N.B.:the fertility rate is not used in the creation of the clusters.
Once the distance matrix is calculated, an algorithm has to be selected for clustering. PAM (partitioning around medoids) is a variation of the k-medoids algorithm, which is a partition-based clustering method. Unlike k-means, which uses means as cluster centers, PAM uses representative objects called medoids as cluster centers. In PAM clustering, the algorithm selects a pre-specified number of medoids from the given dataset. The initial medoids can be randomly chosen or selected using other techniques. The algorithm then iteratively updates the medoids by swapping them with non-medoid objects in order to minimize the dissimilarity (distance) within each cluster. The resulting clusters in PAM are defined by the medoids and the objects assigned to each medoid.
To select the number of clusters, I use the silhouette width. In this case, it suggests 2 as the optimal number of clusters.

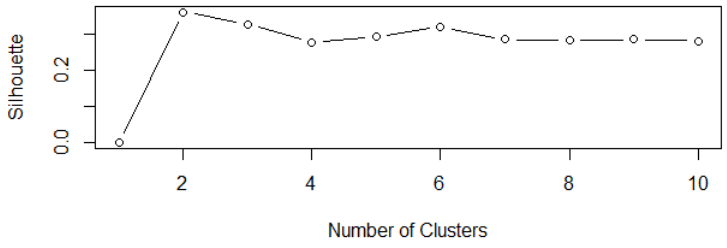


Figure 29: *Silhouette analysis using pam function*

The clusters obtained present these means for all the numerical variables; the `Religion` is the most popular Religion among the countries of the same cluster. It seems that countries which belongs to the first cluster (mostly Islamic African and Asian countries) show a lower means for the variables related to the well being of the countries and a bigger fertility rate. In the second cluster there are all the European Countries.

	`Cluster 1`	`Cluster 2`
`Religion`	Islam	Christianity
`Fertility`	3.20	2.44
`Life.expectancy`	69.18	75.00
`Unemployment`	10.74	8.16
`Alcohol`	0.69	3.31
`Years.school`	6.35	9.40
`Contraceptive`	36.40	58.27
`Maternity.days`	90.12	104.6
`Log_gni`	8.67	9.34
`Freedom`	31.41	68.01
`Migration.rate`	-1.57	-0.16

Cluster 1	Afghanistan, Albania, Algeria, Azerbaijan, Bahrain, Bangladesh, Bhutan, Burkina Faso, Cambodia, Central African Republic, Chad, Comoros, Djibouti, Egypt, Eritrea, Ethiopia, Guinea, Haiti, Indonesia, Iraq, Jordan, Kazakhstan, Kuwait, Kyrgyzstan, Lebanon, Libya, Malaysia, Maldives, Mali, Mauritania, Morocco, Mozambique, Myanmar, Nepal, Niger, Oman, Pakistan, Qatar, Saudi Arabia, Senegal, Sierra Leone, Sudan, Tajikistan, Tunisia, Turkey, United Arab Emirates, Uzbekistan, Yemen
Cluster 2	Angola, Antigua and Barbuda, Argentina, Armenia, Australia, Austria, Barbados, Belarus, Belgium, Belize, Benin, Bosnia and Herzegovina, Botswana, Brazil, Bulgaria, Burundi, Cameroon, Canada, China, Colombia, Costa Rica, Croatia, Cuba, Denmark, Dominica, Dominican Republic, Ecuador, El Salvador, Equatorial Guinea, Estonia, Fiji, Finland, France, Gabon, Georgia, Germany, Ghana, Greece, Grenada, Guatemala, Guyana, Honduras, Hungary, India, Ireland, Italy, Jamaica, Japan, Kenya, Kiribati, Latvia, Lesotho, Liberia, Lithuania, Madagascar, Malawi, Malta, Mauritius, Mexico, Mongolia, Montenegro, Netherlands, New Zealand, Nicaragua, Nigeria, Norway, Panama, Papua New Guinea, Paraguay, Peru, Philippines, Poland, Portugal, Romania, Rwanda, Serbia, Singapore, Slovenia, Solomon Islands, South Africa, Spain, Sri Lanka, Suriname, Sweden, Switzerland, Thailand, Togo, Tonga, Trinidad and Tobago, Uganda, United Kingdom, Uruguay, Vanuatu, Zambia, Zimbabwe

Also in this case the performance of the algorithm can be tested with the silhouette analysis. The average width is 0.36, greater than the one obtained with the k-means algorithm.

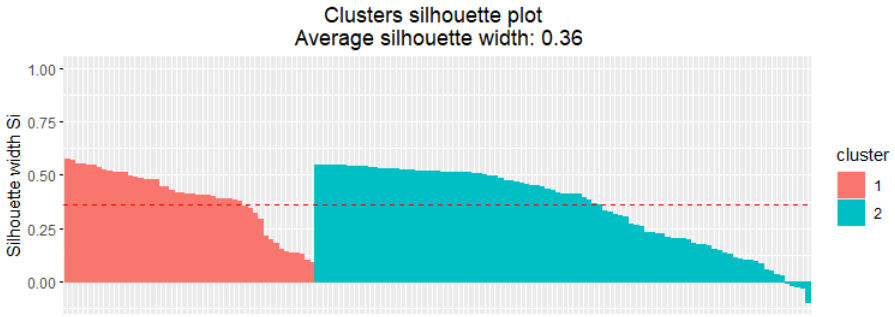


Figure 30: *Silhouette analysis*

4 Conclusions

The paper presents an analysis of the fertility rate among countries. In the first part, which focuses on supervised analysis with the fertility rate as the dependent variable, three different algorithms are employed. The first algorithm introduced is multiple linear regression. All assumptions are tested and most of them are accepted; however, multicollinearity could pose potential issues. With an RMSE of 0.19, multiple linear regression proves to be a highly effective method. To overcome potential multicollinearity problems, a partial least square regression is then conducted, which involves calculating components as substitutes for variables. The optimal number of components chosen is 2, but the resulting RMSE is 0.64, higher than the previous method. The third supervised algorithm used is regression trees, with life expectancy being the most predictive variable, with a threshold of 69 years. A higher life expectancy, as well as GNI, years of school, and maternity days, correspond to a lower fertility rate. The RMSE obtained in this case is 0.70, which decreases to 0.69 when the tree is pruned with three splits, 0.64 using bagging and 0.55 with random forest. Among all the supervised methods, multiple linear regression emerges as the winner.

In the second part, an unsupervised algorithm is presented: clustering. First, the k-means algorithm is applied to the standardized numerical variables. The optimal k value is chosen as 3, and the average width in the silhouette analysis, used to assess performance, is 0.23, which is quite satisfactory. In the first cluster, there are the most developed countries, for which the fertility rate, although not used in the cluster creation, is lower. In the other two clusters, mainly consisting of Asian and African countries, the fertility rate is higher as they are still in the developing stage. To perform clustering while also incorporating the categorical variable `Religion`, the grower distance is used in conjunction with the PAM (Partitioning Around Medoids) algorithm. The resulting clusters amount to 2, and the average width in the silhouette analysis is 0.36. In the first cluster, predominantly African and Asian Islamic countries are found, characterized by a high fertility rate and lower well-being-related variables. In the second cluster, which is predominantly Christian countries, European and American nations are included, exhibiting a lower fertility rate.

References

[1] Clustering mixed data types in r. https://dpmartin42.github.io/posts/r/cluster-mixed-types.

[2] Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani, et al. *An introduction to statistical learning*, volume 112. Springer, 2013.