Università degli Studi di Milano-Bicocca
Corso di Laurea Magistrale in Data Science

A Machine Learning Exploration of Financial Literacy Determinants in Italy

Enrico Mannarino	Thomas Passera	Christian Persico	Giorgia Prina
850859	901685	829558	858740

ACADEMIC YEAR 2022/2023

Contents

Abstract

Financial literacy is remarkably important in today's complex world. Italy is at a disadvantage because its citizens are not as financially literate as those of other major economies. This puts both individuals and the whole country at risk when it comes to financial matters. Using data provided by the answers to a questionnaire administered by the Bank of Italy in 2017, our aim was, in addition to a detailed analysis, the use of machine learning models that allow, by classifying individuals to identify the main factors influencing the level of financial knowledge of Italian adults. This work is intended as a tool to support national and international institutions, as well as policy makers and financial analysts who, on the basis of the results obtained, have the possibility to intervene in a targeted manner to contribute to the development of financial education in the country.

"A person can either discipline their finances
or their finances discipline them"

O. Woodward

Introduction

Financial literacy is an indispensable tool in navigating today's complex economic landscape. Yet, Italy finds itself in a precarious position, with lower financial literacy levels compared to its G20 peers. Only 37 percent of adult Italians grasp basic financial concepts, compared to the EU's 52 percent average (S&P Global FinLit Survey). This discrepancy not only endangers individual financial stability but also threatens Italy's economic resilience. The ramifications of this disparity are far-reaching. Financially illiterate individuals are more prone to making poor financial decisions, leading to inefficient resource allocation and heightened vulnerability to economic shocks. This paper delves into Italy's financial literacy challenge, identifying the most vulnerable groups within the population. Our reasoning begins with the publication of the first OECD report about financial literacy in 2017. Starting in 2016, the OECD began conducting its first financial literacy surveys, analysing data from 30 countries and economies and using instruments that could be internationally comparable. The results, in line with those above-mentioned of the S&P Global FinLit Survey, revealed serious gaps in global financial literacy, so much so that there was an urgent need to develop long-term strategies to fill these gaps, as well as periodic measurements to monitor the progress of these strategies or remaining weaknesses. The final report compiled in 2017 by the OECD [1] presents data from 21 countries, based on interviews conducted with a sample of 101,596 individuals aged between 18 and 79. Each country followed the guidelines and used the questionnaire produced by the INFE so that the data could be compared with each other. The result was an OECD-INFE harmonised questionnaire which was used to analyse and compare three financial literacy areas: knowledge, behaviour and attitudes. To this end, the OECD has developed a particular methodology based on the construction of indices to measure the categories listed above and which we use to get a comparable result but with a focus on Italy. Next we will explain the construction process in more detail but for now we are interested in the overall index, a sum up of the 3 above. It consists of 7 points relating to financial knowledge, 9 points relating to financial behaviour and 5 points relating to financial attitudes. The maximum possible score is 21 points, with the average of the 21 countries considered standing at 12.7 points. In this context, Italy is among the four countries with the lowest scores of less than 12 points. Knowing now the Italian position with respect to the other 20 countries analysed, we can better delve into the analysis of this specific case. Given these premises, our study aims to identify the main factors influencing the level of financial knowledge of Italians aged 18-79. In other words, our approach to the issue involves producing a pure analysis of the available dataset, which aims to discover the most pressing issues in the Italian economic literacy landscape. To do this we will use some machine learning models that are able to classify individuals on the basis of the influential factors found. More precisely, we will use Logistic regression, Decision Tree, Random Forest and Gradient Boosting in accordance with the study by S. Levantesi and G. Zacchia [2] in which they evaluate their accuracy and conclude that they can be a good complement to traditional models. Our intent was to start from this to produce increasingly accurate results using similar models in order to provide a useful tool for policy makers and institutions, who will have the right knowledge to propose tailor-made solutions and sensible reforms for specific segments of the population with regard to financial literacy. In any case, a general interest and concern arose from the 2017 results. Therefore some reforms have already been implemented in some states. One of the most set goals was a long-term one to introduce and improve financial education in schools in order to cultivate financially aware citizens capable of effectively managing their assets, especially during crises such as the COVID-19 pandemic. The OECD believes that all the data from the questionnaires that will be submitted to population samples in the years to come will certainly even serve as a test to check the effectiveness of reforms similar to the one just mentioned - always taking contingent factors into account. However, for the time being, it seems to be too early to make such assumptions. Furthermore, while the importance of imparting financial education to the younger generation is widely recognised, it is equally essential to consider the existing adult population, particularly from the perspective of public policy and voters. Neglecting the financial literacy of the existing population could have negative consequences. In conclusion, we can sum up the OECD's thinking and the ideological heart of our work with a sentence describing the importance of financial literacy, namely: *"financial literacy, including awareness, knowledge, skills, attitude and behaviour, is crucial for making wise financial decisions and achieving individual financial well-being"*.

1. Dataset and data preparation

Since 2017, the Bank of Italy has been conducting a triennal survey on financial literacy in Italy [4] and the subject of this work will be the anonymously distributed 2017 data. The survey in Italy was submitted to a sample of 2,376 people with two distinct methods: 60% of indivisuals responded via a tablet device designed to be easily used by all subgroups of the population (even the less educated or the elderly), while the remaining were interviewed personally using CAPI methodology (Computer Assisted Personal Interviews).

The dataset consists of the answers collected from the 21-question questionnaire comprising both single and multiple choice questions covering socio-demographic and quantitative information. As a first step, we renamed some variables in order to make them more understandable during the analysis. In addition, during the data cleaning process we checked that there was no need to correct or deal with null values present, as these are filter questions of the questionnaire (depending on the answer given, the respondent can be directed to specific sections of the questionnaire or can be excluded from certain questions or sections, thus saving time and avoiding collecting irrelevant data).

1.1 Financial literacy scores

Similar to the approach adopted by the OECD, we have chosen to categorise some of these questions into four distinct indices:

- Financial knowledge score (0 - 7)
- Financial behaviour score (0 - 9)
- Financial attitudes score (1 - 5)
- Overall financial literacy score (1 - 21)

The knowledge component assesses the understanding of basic concepts, a prerequisite for making sound financial decisions. The main topics of the assessment are three and represent the standard in the financial literature [3]: understanding simple and compound interest, inflation and the benefits of portfolio diversification. The literature states that high levels of financial knowledge are associated with higher participation in stock markets and a reduction in debt accumulation, as individuals who attain this rating are able to make informed decisions and handle periods of crisis more intelligently.

The behavioural component measures how common are the behaviours that often indicate a greater ability to adequately manage financial resources in the sampled population. This component is the one that really shapes people's financial situation. This index is calculated by considering questions that assess whether people are able to manage their own or their family's financial resources, whether they are able to pay their debts or utilities without worrying, and whether they inform themselves before making investments. A particular focus is placed on active savings. The OECD also highlights how the increasing digitalisation of finance is altering consumers' interactions with a wide range of (new) financial providers. This has resulted in new trends (trends/modes) that have altered and are still altering users' investment behaviour.

The attitudes component assesses individual characteristics such as preferences, beliefs, and non-cognitive skills, all of which can have an impact on personal well-being. Specifically, this component seeks to measure an individual's inclination to allocate a substantial portion of their income towards consumption, irrespective of their future financial needs. Even if an individual has basic financial knowledge, he or she will still be influenced by factors such as beliefs or preferences when making a financial decision (in short, this is a possible irrational behaviour).

The overall financial score summarises the three indicators into a single score, which is the sum of the three previous components. Interestingly, the component with the greatest weight in this score is financial behaviour, contributing no less than 9 of the total possible 21 points. It is clear from this that positive financial behaviour has a significant impact on overall financial well-being. Currently, Italy has a score of 11, which is below the average of the G20 countries, which is 12.7.

1.2 Additional financial variables

Not all answers to the questionnaire were taken into account in the creation of the indices, but we have nevertheless included the remaining ones in the analysis, as they can potentially provide other useful information for our purposes. In particular, we have named the response variable to the question QF4 as *risk capacity* because it measures respondents' answers to the following question: "*What if you were to incur a significant expense today, equivalent to your monthly income, would you be able to pay for it without borrowing or asking for help from family or friends?*" This gives us an indication of the respondents' financial ability to manage significant expenses without resorting to loans or outside assistance.

Closely linked to this question are QF11 and QF13. The former was considered as *financial fragility* as it asks respondents whether they had difficulties covering their living costs with their income in the last 12 months. This question provides information on respondents' financial resilience in the face of daily expenses. The second question, on the other hand, measures *financial robustness* as it asks respondents how long they could get by if they lost their main source of income. This helps us assess people's financial fragility and their ability to cope with emergency situations or a sudden loss of income.

Finally, responses were also considered to analyse the financial activity of Italians, in particular which financial products they have bought in the last two years and whether they still own them (Qprod1c - *buy financial products*), and to obtain information on the subject of pensions (QF9 - *pension fund*), a topic of great relevance in Italy.

2. Comparative analysis of socio-demographic factors and indices

After creating the indices, we proceeded with the analysis of the renewed dataset containing the 19 variables of interest among which we find, in addition to the indicators mentioned above and the extra questions we have kept, socio-demographic information such as age and education level that we have grouped into intervals - *age* and *education group* - in order to simplify the analysis and allow more direct comparisons. Specifically, the age ranges created are 18-30, 31-50, 51-70 and 71+, while the education level groups are 'primary or less' (if the individual has completely or partially finished primary school, or has no education at all), 'secondary' (if the individual has completely or partially finished secondary school) and 'university' (if the individual has a university degree). A final step of feature extraction was performed for classification purposes, creating a new binary variable - financial literacy grade - which approximates the respondents' degree of financial knowledge in two classes, below and above sufficiency respectively (equal to 4 out of 7 points).

2.1 Initial overview

2.1.1 Composition of the processed dataset

With reference to the Figure 2.1 presenting the distributions of the main variables, we note that in the Italian sample under examination, the selection of participants constitutes a good representation of the different geographical areas, with the majority coming from the northwest (26.8%) and a minority from the islands (11.4%). The same observation applies to the gender breakdown, where the number of men (1166) and women (1210) nearly equals each other. Moreover, almost all the subjects in the sample were born in Italy, only 2.6% abroad. On average, each household surveyed is made up of between 3 and 4 members, with some exceptions of households with 1, 5, or 6 members; a similar trend was recorded for age groups, in which the majority of respondents (more than 50%) fall into the middle age groups of 31-50 and 51-70 years, while the remainder are in the 18-30 and 71 or more age groups.

More unbalanced are the degrees of education, in fact 76.4% of the respondents have at least attended secondary school, 22.5% have a university level of education, and a very small part of 1.1% have a primary school leaving certificate or not even that. Regarding the type of employment, we note that the majority (60.4%) are employees and pensioners. Having made this brief overview of the socio-demographic components, let us turn to the scores and the relevant questions.

2.1.2 Distribution of the variables

The financial knowledge score has an approximately Normal trend, with a peak close to sufficiency; in fact, 53.7% of the respondents have a score above sufficiency (at least 4 out of 7). A similar conclusion can be deduced for the financial behaviour score, with a more symmetrical trend in which we attest that only 36.2% of individuals reach sufficiency, confirming what the OECD and Bank of Italy documents indicate. It would also seem that just over half of people (52.3%) do not feel able to meet a significant expense equal to their monthly income today without borrowing. This demonstrates a weak financial situation within the sample, highlighting a significant challenge for many participants in managing unexpected expenses without resorting to external financial loans or assistance. An equally worrying result emerges when we consider financial fragility, as a considerable percentage of participants admitted to having difficulty meeting their financial obligations using only their income in recent months. This underlines the need to examine strategies to improve financial resilience among participants. Given that the majority of respondents recognised difficulties in meeting their recent financial commitments, it is reasonable to expect that they would not be confident of being able to sustain a long period of financial inactivity should they lose their main source of income today. Indeed, this consideration is supported by the data. Analysing the financial robustness, we note that the vast majority (74.1%) stated that they would not be able to survive for more than three months if they were to lose their source of income today. Continuing the analysis with a focus on financial activity, more than 950 individuals bought a financial product in the two years prior to completing the questionnaire and still own it. However, it should be noted that the number of individuals active in this area is not particularly high, as the majority of respondents have not made any purchases or no longer own financial products

(59.6%). To conclude, let us take a quick look at the topic of pensions, where the graph shows that only a small percentage of participants opted to join a private pension system (10.3%). This suggests that most individuals in the sample prefer to rely on the public pension system, and could be an important indicator for assessing the population's confidence or financial choices regarding their pension future.

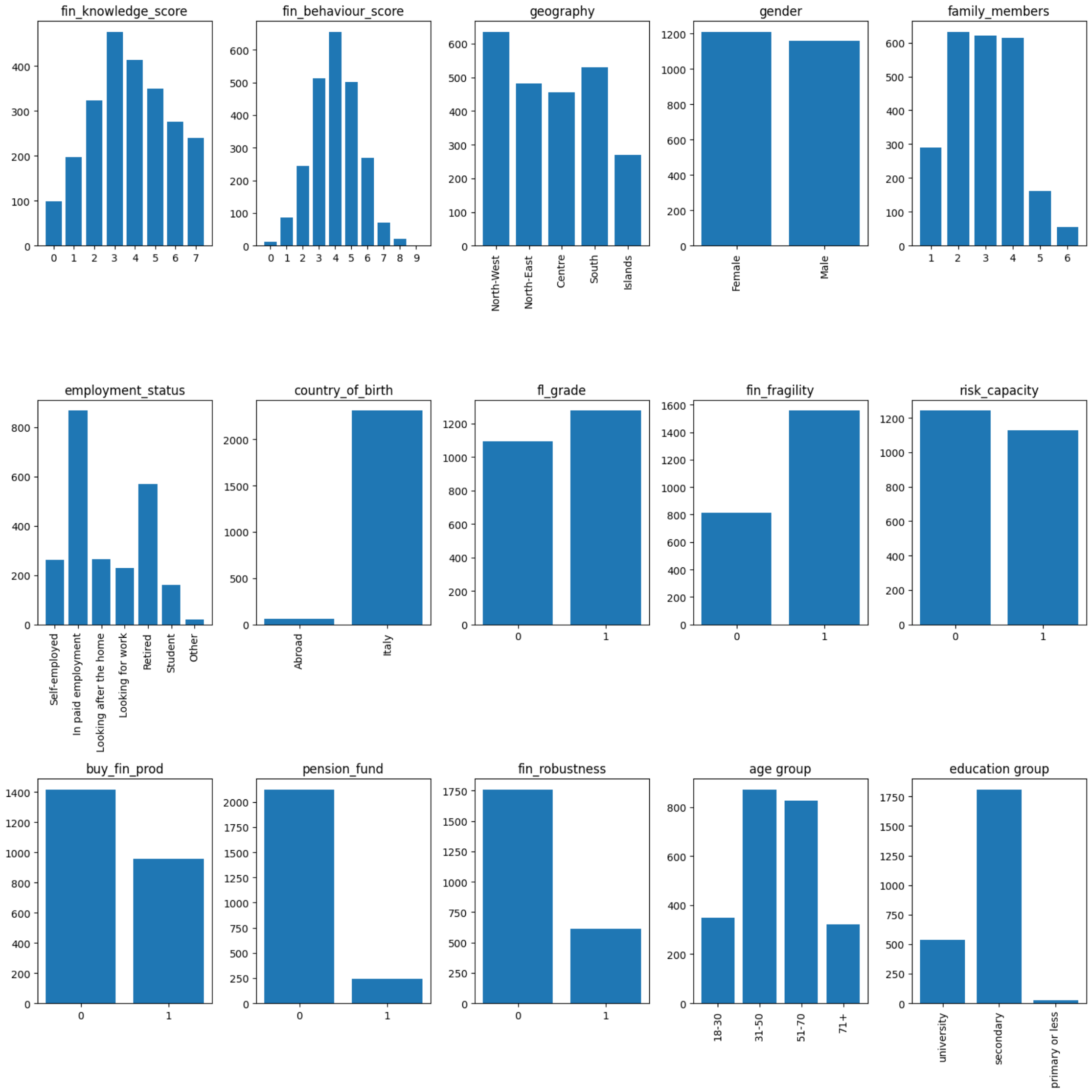


Figure 2.1: Distribution of features

With regard to the correlation analysis of the continuous data at our disposal, no significant links emerged between the variables except between the scores and the overall financial literacy score, where the highest correlation was 0.8 between the financial knowledge score and the overall financial literacy score.

2.2 Relationship between financial knowledge and other variables

In order to obtain a clearer view of the relationships between the qualitative variables, let us consider below the bar charts in the Figure 2.2 and 2.3.

2.2.1 Age

It is evident from the first graph that the presence of poorly educated individuals is particularly rare and occurs mainly in the older age brackets, particularly in the 71+ age group. In contrast, the age groups that include younger individuals, i.e. 18-30 and 31-50, show a majority of people with a high school diploma and a significant percentage of university graduates.

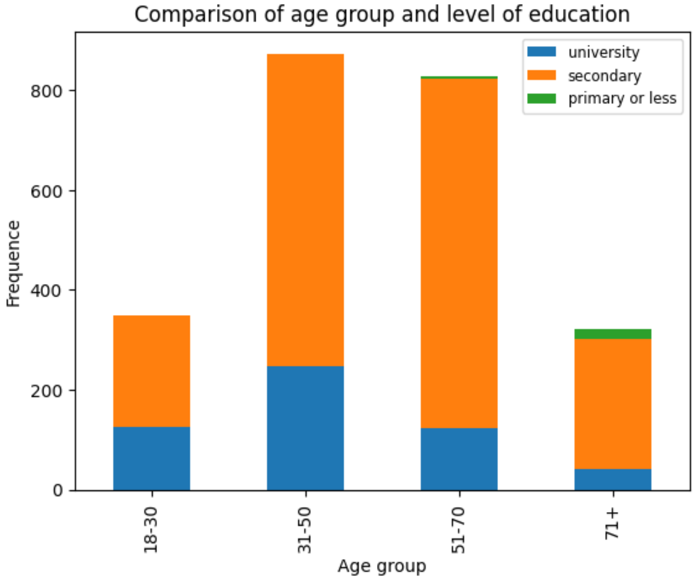


Figure 2.2: Comparison of age group and level of education

It is clear that the level of financial knowledge is not uniform across the population, and to get a clearer picture we refer to the first graph of 2.3, where we first examine the relationship between this and age. Initially, a balance seems to emerge, with an apparently even distribution of above and below average subjects for each age group. However, a more detailed analysis reveals that in the intermediate age groups there is a majority of individuals with high scores compared to the age group of 71 and above. Referring back to 2.2, we note that it is precisely those in the 18-30 and 31-50 age groups who obtain higher scores, those who appear to have had a higher level of education.

2.2.2 Education

This relationship is highlighted even more clearly by the second graph in 2.3, which compares the link between financial knowledge and education level. Indeed, the Bank of Italy points out that education is one of the most important factors in ensuring adequate levels of understanding of financial concepts, stating that in Italy: "The average knowledge score drops from around 4 for university graduates to around 3.2 for those with secondary education and to 2 for those with a lower level of education". This is confirmed in our study in which we note that subjects with a primary school licence or less scored low compared to those with a secondary school or university degree. Analysing the values for secondary school and university levels individually, we note that, in the first case, about half of the respondents in this bracket obtained above-average scores and the other half below-average scores; whereas in the second case, there is a propensity to obtain higher scores, with more than half of the respondents found to have obtained above-average scores.

2.2.3 Gender

The OECD's report on financial literacy also devotes a few paragraphs to in-depth studies on categories of interest. For example, it mentions the importance of financial literacy for women and young people. The main reason is that women, statistically, have longer lives, shorter careers and lower earnings than men on average in all countries, including Italy. Therefore, improving women's financial literacy is essential to promote their long-term economic empowerment. This is why this is of particular relevance for policies adopted by institutions and includes a specific analysis of differences in financial literacy levels between genders. In fact, from the third graph of 2.3, we can see that the results are in line with those obtained from the studies: there is a slight difference between men and women. On average, female respondents scored about 50% below average and the remaining 50% above average. As far as men are concerned, the balance shifts in favour of higher scores. Although not significantly, there is still a difference between the sexes. The OECD provides its own note on this, noting that in Italy there are indeed gender gaps in financial literacy, although smaller than in other countries. More specifically, women with a high level of education have lower financial knowledge scores than their male peers.

2.2.4 Geography

The last graph in Figure 2.3, shows the comparison between financial knowledge and geographical area. In general, the respondents in each region seem to be evenly divided between those who scored high and those who scored low, with a slight exception in the North-East, where there seems to be a slight majority of users with higher ratings.

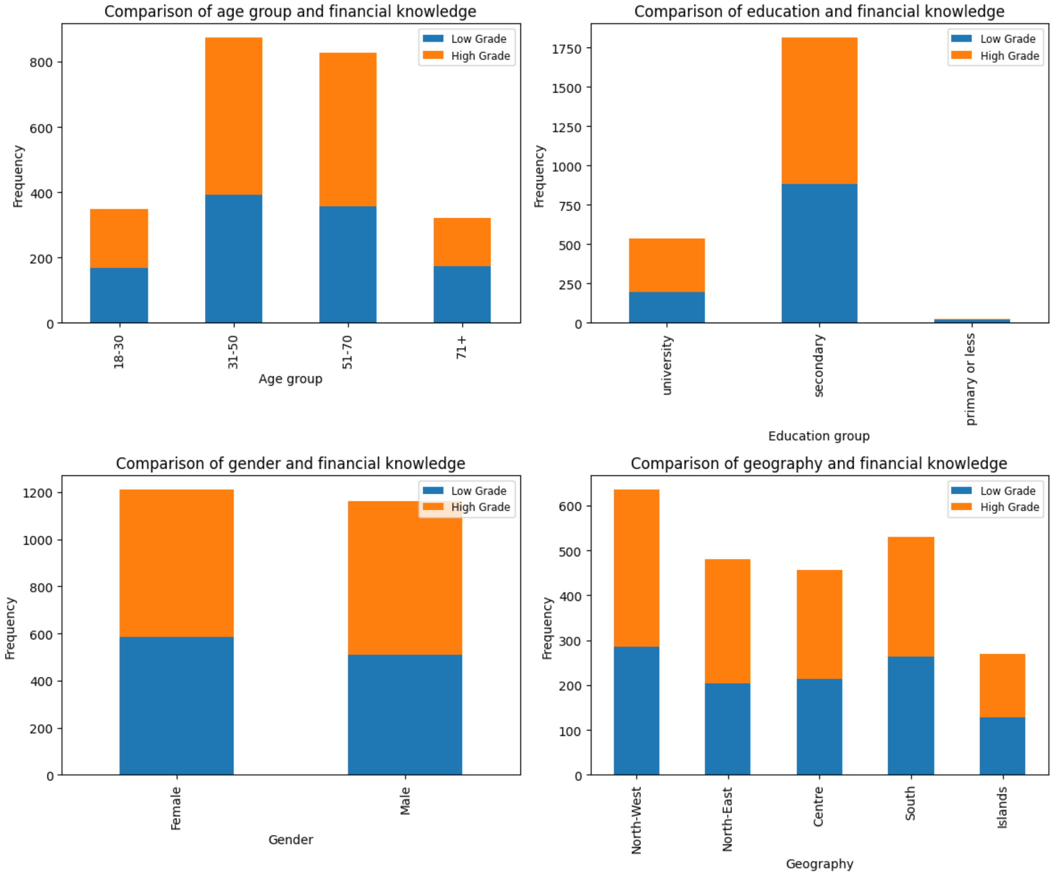


Figure 2.3: Comparison of socio-demographic factors and financial knowledge

2.3 Aggregating variables searching for insights

In conclusion, we propose the summary table 2.1 of the aggregate mean scores gender, age and degree of education, i.e. the variables that, according to the literature, have the greatest impact on financial knowledge and behaviour. Overall, the highest scores recorded for each score are as follows: 4.9 for financial knowledge, 4.6 for financial behaviour and 4.667 for financial attitude. It is noted that, for financial knowledge and financial behaviour, the highest scores were obtained by male subjects with a university education level. On the other hand, for financial attitude, the highest scores were achieved by female subjects with primary level education or less. We also note that the lowest scores are at 1.5 for financial knowledge, 2.4 for financial behaviour and 2 for financial attitude. In this case, for financial knowledge and financial behaviour, the lowest scores are associated with male individuals with primary level education or less. In contrast, for financial attitude, the least satisfactory results concern female participants with primary education.

Let us now proceed with an in-depth analysis of the variables under consideration. With regard to gender, we note that in most cases male users report higher average scores than their female counterparts, especially in the two scores with the greatest influence on the overall score, namely financial knowledge and financial behaviour. According to the OECD, women with a high level of education have lower financial knowledge scores than their male peers. With regard to age, we observe a general trend in which users aged between 51 and 70 tend to have higher scores. Furthermore, in relation to the education variable, it is clear that users with university education levels tend to obtain higher scores. It should be noted that, for example, the association between a higher level of education in older subjects could reflect greater accumulated experience than in younger subjects.

It is to be understood that the numerosity of each group varies and, for some, is even zero (e.g. men aged between 18 and 50 with minimal education).

Table 2.1: Financial literacy scores by socio-demographic class

Sex	Men											
Education	Primary or less				Secondary				University			
Age	18-30	31-50	51-70	71+	18-30	31-50	51-70	71+	18-30	31-50	51-70	71+
Knowledge	-	-	1.5	2.56	3.66	3.69	3.89	3.55	4.13	4.67	4.9	4.75
Behaviour	-	-	3.5	2.44	3.59	4.05	4.12	3.8	3.67	4.25	4.49	4.6
Attitudes	-	-	2.5	3.22	2.64	3.03	3.17	3.18	2.79	3.17	3.46	3.65

Sex	Women											
Education	Primary or less				Secondary				University			
Age	18-30	31-50	51-70	71+	18-30	31-50	51-70	71+	18-30	31-50	51-70	71+
Knowledge	2	-	3	2.6	3.37	3.66	3.79	3.21	3.82	4.01	4.25	3.35
Behaviour	5	-	4	3.2	3.55	4.04	4.07	3.96	3.84	4.24	3.92	4.4
Attitudes	2	-	4.67	3.23	2.8	3.08	3.29	3.2	3.12	3.18	3.13	3.31

The results described are consistent with the findings of the Bank of Italy and the OECD and are in line with existing economic statistics. However, it is crucial to consider that each country has specificities linked to its history and current demographic composition. Therefore, it is essential to conduct studies that take these aspects into account before drawing definitive conclusions.

3. Machine learning models

In order to answer the research question and find out which factors influence the financial knowledge of the considered sample, we leveraged the inherent opportunities offered by machine learning models found in the literature.

Indeed, through a classification model we can assess how much the independent variables influence the assignment to one of the categories of the dependent variable. Specifically, by setting the financial knowledge level - *fl grade* - as a binary response variable, we can determine which independent variables are more important in correctly classifying the financial knowledge of the examined individual and which, regardless of their values, do not appear to influence the model's decision.

After ensuring that we do not have imbalanced classes - 46% of individuals are below the sufficiency, while the remaining 54% are above - we partitioned the dataset into a training set (80% of the data) and a test set (20% of the data). Given the dataset's limited size, a 10-fold cross-validation procedure was implemented to better compare the models' performances.

Below, in the first part, we provide a brief theoretical description of the models used to support our research. In the second part, we present the results.

3.1 Models presentation

3.1.1 Logistic

As a starting point, we have chosen a fast and computationally inexpensive model suitable for the problem, namely logistic regression. This is a technique used to analyse data between a binary response variable and one or more independent variables that can take any form. The model in fact expresses the probability that, for a given record, the dependent variable belongs to a certain class (0 or 1).

$$P(Y=1) = \frac{1}{1+e^{-(\beta_0+\beta_1 X_1+\beta_2 X_2+...+\beta_p X_p)}} \tag{3.1}$$

The goal is to accurately calculate the values of the estimated coefficients β so that the model can optimally explain the data used for training. Subsequently, we increased the complexity of our model with a view to improving the metrics, with the aim of making the identification of the determining independent variables more reliable.

3.1.2 Decision tree

A decision tree model consists of a non-parametric supervised learning method. This model is structured as a tree in which each node represents a decision related to a variable. The branches branching off from each node indicate the possible outcomes of that decision and lead to further decisions. The leaves constitute the final nodes of the decision path where there are no further branches and contain the class predictions associated with that particular decision path.

$$\hat{f}_{DT}(x) = \sum_{j \in J} \hat{Y} R_j \mathbb{1}_{\{x \in R_j\}} \tag{3.2}$$

where $\hat{f}_{DT}(x)$ represents the prediction $\hat{f}$ of the model for input x and $\sum_{j \in J}$ denotes the sum over all sets of terminal nodes R_j in the tree. Moreover $\hat{Y}$ represents the prediction of the j-th terminal node while $\mathbb{1}_{\{x \in R_j\}}$ is the indicator function that takes values 1 or 0 depending on the condition that x belongs to R_j.

The size and complexity of the tree is controlled by a stopping criterion - *ccp alpha* - necessary to penalise the complexity of the decision tree. Higher values of CCP lead to greater simplification of the tree, removing branches that would only contribute a small amount to reducing training error. The value we chose was 0.002, as a compromise to maintain a balance between complexity and generalisation capacity. The dataset, in fact, contains a large number of variables but a small number of observations, so a too high CCP would have oversimplified the tree and would not have allowed the model to capture complex relationships between the data; conversely, a too low CCP would have maintained a very complex tree and likely overfitting.

3.1.3 Random forest

It is a model that combines the predictions of several decision trees to improve overall performance and reduce the risk of overfitting. Specifically: the model trains trees on random subsets of the data and independent variables. When making predictions, it consults the trees and returns the most frequent class predicted by them.

$$\hat{f}_{RF}(x) = \frac{1}{B}\sum_{b=1}^{B} \hat{f}_{DT}(x|b) \tag{3.3}$$

where $\hat{f}_{RF}(x)$ represents the prediction of the model and B represents the total number of decision trees. Indeed $\hat{f}_{DT}(x|b)$ represents the prediction of the single decision tree DT for input x while $\sum_{b=1}^{B}$ is the sum of the predictions of the decision trees successively divided by their numerosity. Commonly, values greater than 500 trees are used to give the model greater robustness; however, in this specific situation, we opted for 300 trees in order to achieve good generalisation capability without compromising performance, as more trees did not lead to significant improvements in the metrics. Furthermore, we agreed that it was necessary to set a minimum number of 3 nodes per decision tree in order to allow the model to also consider variables of lesser importance that might otherwise have been ignored.

3.1.4 Gradient boosting

Lastly, let us consider the Gradient Boosting model, an evolution of the previous cases that combines several models sequentially to achieve better performance, in this case weak (small) decision trees. Such a model, being more complex, may be able to capture more intricate relationships between data that may have gone unnoticed so far.

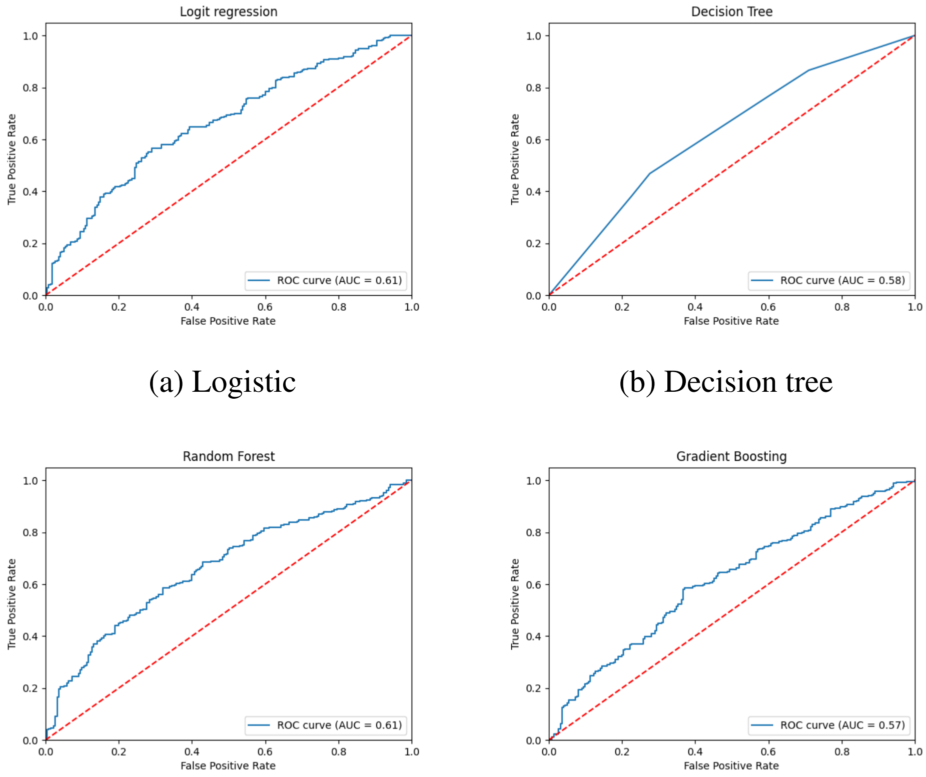
$$F_m(x) = F_{m-1}(x) + \lambda \cdot \gamma_m \cdot h_m(x) \tag{3.4}$$

where $F_m(x)$ represents the prediction of model m given an input x and $F_{m-1}(x)$ represents the prediction of model $m-1$ given an input x. λ defines the learning rate of model m with respect to $m-1$, γ_m represents the weight of the model and $h_m(x)$ with respect to the residual error. Finally, $h_m(x)$ is the prediction of the weak decision tree m for input x.

A substantial number of trees of 500 was adopted, even more than those considered so far. The learning rate, of 0.1, allows the model to achieve better performance by iteratively minimising the error.

3.2 Metrics of the models

In general, the performance obtained on each model is rather limited. The accuracy value on the test set ranges from a minimum of 0.58 obtained by the gradient boosting to a maximum of 0.61 obtained by both logit and random forest. Similar results concern the AUC, the area under the ROC curve (Figure 3.1), which indicates the rate of true positives (sensitivity) positioned on the y-axis, compared to false positives (specificity) positioned on the x-axis, of the classification.



(a) Logistic

(b) Decision tree

(c) Random forest

(d) Gradient boosting

Figure 3.1: ROC curves for ML models

Again, the metrics range from a minimum of 0.57 obtained by the Gradient Boosting to a maximum of 0.61 obtained by both the Logit and Random Forest. We can therefore conclude that despite the computational and structural differences of the models, the results do not differ significantly from each other. It is interesting to note that the only parametric model (logit) produced results exactly equivalent to those of the Random Forest model, the best of the non-parametric models.

3.3 Feature importance

Having defined and estimated the logistic model, the variables with which the most significant estimated coefficients are associated are: *financial robustness* (0.5107), *risk capacity* (0.4185), *pension fund* (whether he/she bought a private pension plan)(0.2941), *country of birth* (0.2407), *education* (-0.1959). These values show that individuals in our sample who do not have short-term economic difficulties and protect their future consumption by giving up part of their present consumption, i.e. by saving or investing, perform better on the financial knowledge questions.

The graph in Figure 3.2 shows that the tree made only three decisions in relation to the features *risk capacity*, *education* and *financial robustness*, with feature importance values of 0.51, 0.39 and 0.11 respectively, as shown in Figure 3.3.

Interestingly, it is precisely these that were of considerable importance in the previous regression, and the new results confirm this significance. We can therefore deduce that both models produce similar conclusions, i.e. that the individuals examined who possess a good level of education and who are not in an economically difficult situation appear to have a deeper understanding of the basic concepts of economics.

The results obtained for the tree-based ML algorithm (3.2) represent the best decision tree for the FL data used. We see that the best tree has 5 terminal nodes (4 splits) and the root node splits on risk capacity ≤ 0.5 into true and false. Each node shows the expected class (1 or 0) and the percentage of observations in the node.

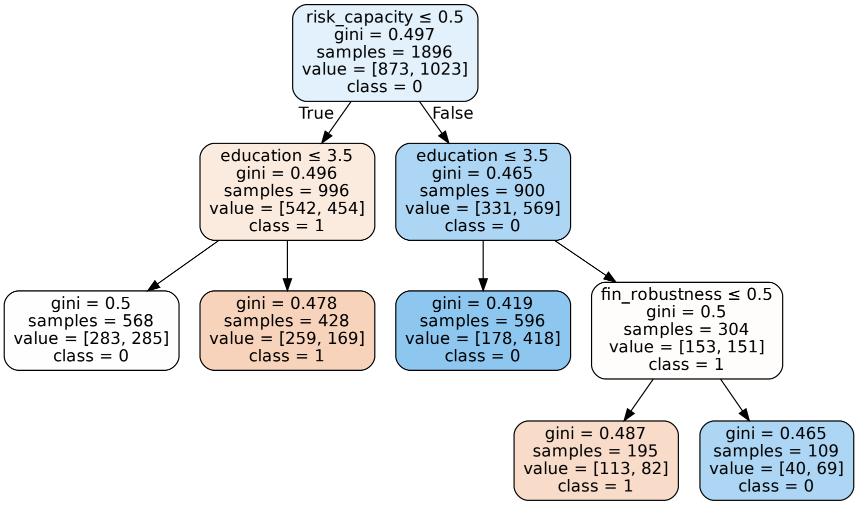
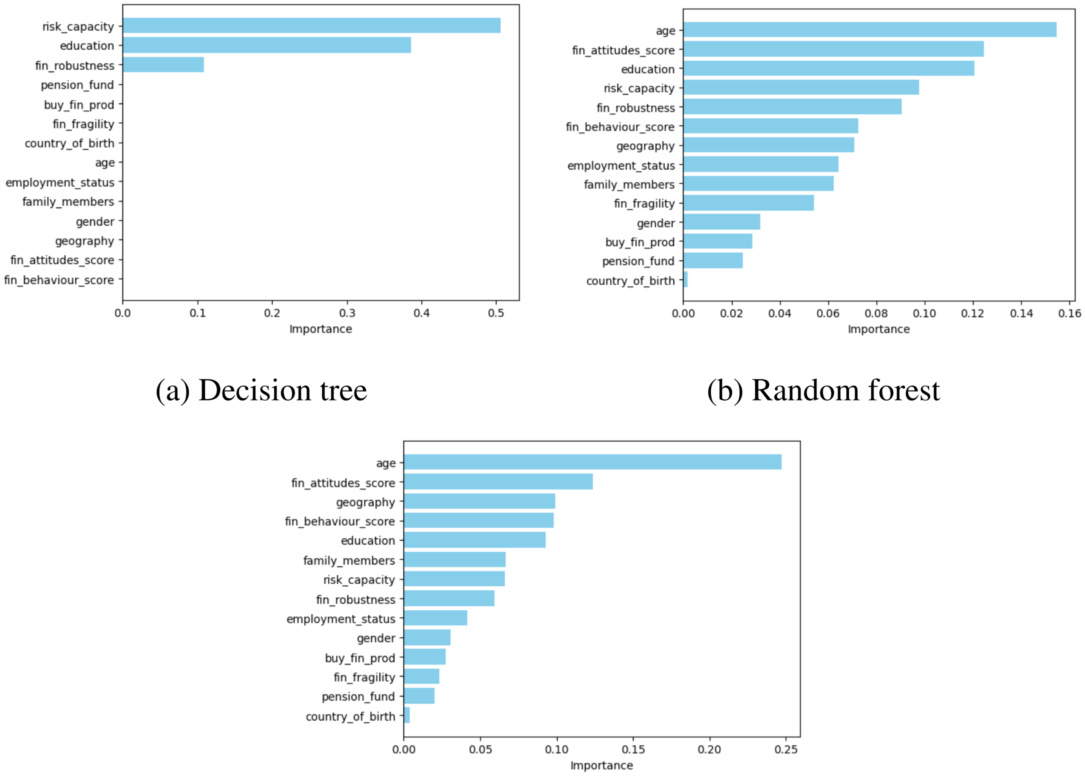


Figure 3.2: Decision tree for sufficient financial literacy

The large number of features makes it disadvantageous to use a tree model, which has been shown to underperform, if only slightly, simple logistic regression. This drawback is due to the use of locally optimal solutions that may not be able to guarantee globally optimal trees. It was therefore deemed necessary to adopt a model of greater complexity, with the aim of improving the predictions made by the model. We thus opted for a Random Forest.

Through the use of the Gini Impurity Index, it is possible to determine which variables are most frequently used by the trees for the subdivision of the sample. The two features considered by the model as most relevant are *age* (0.1546) and *financial attitudes* (0.1246), as we can see in Figure 3.3. Followed by the same 3 variables - *education* (0.1206), *risk capacity* (0.0977), *financial robustness* (0.0905) - already used in the previous decision tree. Starting from this model, the variables indicating the household's financial robustness give way to variables more related to the individual person and not to his or her surroundings. The respondent's age and his personal characteristics, preferences, beliefs and non-cognitive skills rise to the top as incisive determinants of his financial knowledge.

The last model allowed us more room for manoeuvre in configuring the parameters. However, more freedom also means more complexity in the implementation. Furthermore, gradient boosting, compared to random forest, is more sensitive to noisy data due to the different method used to update the parameters. This greater complexity did not lead to a significant improvement in performance in terms of accuracy, however it did confirm, at least in part, the results obtained by the random forest. In particular, the two determining variables remained the same as in the previous model, namely: *age* (0.2471) and *financial attitudes* (0.1237). Next we have the features *geography* (0.0992), *financial behaviour* (0.0980) and *education* (0.0929) - as shown in Figure 3.3.



(a) Decision tree

(b) Random forest

(c) Gradient boosting

Figure 3.3: Feature importance with ML models

3.4 Results comparison

The results obtained by comparing the models are very similar, so much so as to minimise the need for additional computing capacity and memory to be devoted to models more complex than an ordinary logistic regression. It should be emphasised that the data from the questionnaire are difficult for a model to read due to the presence of latent variables, such as the income of the individuals and the sector in which they are or were professionals, which hinders an effective understanding of the relationships between the variables in the dataset. Despite this premise, it is essential to specify that the results obtained are consistent with the available literature on financial literacy. The present analysis therefore aims to support these conclusions, offering a useful tool for policy makers to obtain an overview of the Italian situation in the first year in which the questionnaire was administered.

Conclusions

The aim was to offer a valid tool to support the arguments put forward in the literature regarding the financial literacy of Italians. By analysing the parameters estimated by a statistical classification model, it is possible to determine which factors most influence the decision to assign a given class to the record under examination. The context in which this work will be carried out is, on the other hand, a socio-political context that has set itself the goal of improving the well-being of people and in particular of the Italian population.

The importance of these results lies precisely in their applicability. Each model either reinforces the results of the previous one or raises interesting questions about the relevance of certain variables. By subjecting this information to four different classification models, each characterised by an increasing degree of complexity, we were able to identify the decisive common features among them.

Within our sample, what emerges clearly is that, in the first place, the level of education is the main component, followed by features concerning the economic stability of the household. Finally, we would like to mention the age factor, which, although it plays a less important role than the previous ones, cannot be underestimated. We have noticed that in general people feel economically uncertain and dissatisfied with their economic situation or income. They feel that they are unable to cope with a sudden termination of an employment relationship, as well as unable to handle a large unexpected payment in a short time without resorting to loans of any kind. A general malaise is perceived from this point of view. Inability indicates a lack of clarity in knowledge and, consequently, in behaviour. For this reason, we often find financial behaviour and financial attitude as determining factors in our analyses, in addition to the degree of education, which is one of the main discriminants in our models.

Our results are consistent with the existing literature, and the use of models has proved crucial in uncovering subtle and deep relationships that would otherwise have escaped observation. Increasingly implementing such models will allow us to obtain results of increasing precision, an imperative need for policy makers and institutions seeking to formulate more targeted solutions.

References

[1] *G20/OECD INFE Report on Adult Financial Literacy in G20 Countries*. OECD, 2017. URL: https://books.google.it/books?id=RJ_RxQEACAAJ.

[2] Susanna Levantesi and Giulia Zacchia. “Machine Learning and Financial Literacy: An Exploration of Factors Influencing Financial Knowledge in Italy”. In: *Journal of Risk and Financial Management* 14.3 (2021). ISSN: 1911-8074. DOI: 10.3390/jrfm14030120. URL: https://www.mdpi.com/1911-8074/14/3/120.

[3] Annamaria Lusardi and Olivia S Mitchell. “Financial literacy around the world: an overview”. In: *Journal of pension economics & finance* 10.4 (2011), pp. 497–508.

[4] Antonietta di Salvatore et al. *Questioni di Economia e Finanza*. Banca d’Italia, 2018.