Capstone Project Final Report

Marisa Donnelly

Student number: 501172691

Supervisor: Tamer Abdou

Submission date: July 29, 2025

Toronto Metropolitan University

Table of Contents

Introduction

Financial literacy significantly influences people's ability to manage money, plan for the future, and navigate formal financial systems. This capstone project explores whether machine learning models can classify countries or regions as having high or low financial literacy using the World Bank's Global Findex dataset (Demirgüç-Kunt et al., 2018). The objective is not only predictive accuracy, but also transparency and interpretability, which are essential when socio-demographic features are involved.

Previous studies have emphasized the importance of financial literacy in shaping individual and societal economic outcomes (Lusardi & Mitchell, 2014; Grohmann et al., 2018). As financial markets become more complex, financially literate populations are better positioned to avoid debt traps, utilize banking services, and achieve long-term financial well-being (Fernandes et al., 2014; Guiso et al., 2015). Machine learning has emerged as a promising approach to capture these patterns and assess risk factors efficiently (Yue & Zhu, 2025; Lokanan et al., 2021).

This project addresses two research questions:

1. Can behavioural financial indicators from the Global Findex dataset be used to classify financial literacy levels across countries?
2. Which machine learning model best balances predictive accuracy and interpretability for this task?

Data Preparation

The dataset is based on the 2021 Global Findex survey. Data cleaning involved removing columns with 100% missing values and replacing partial missing values with column-wise means. A binary target variable was created from a financial literacy proxy score, which averaged six key behavioural indicators (e.g., account ownership, savings, emergency funds access). These features were scaled appropriately or encoded depending on type.

Key transformation steps:

- Combined cleaning, encoding, and scaling in a scikit-learn Pipeline using ColumnTransformer
- Applied transformations **only to training data**, then transformed test data with fitted parameters
- Handled categorical encoding with OneHotEncoder (handle_unknown='ignore') and scaling with StandardScaler

Exploratory Visualization

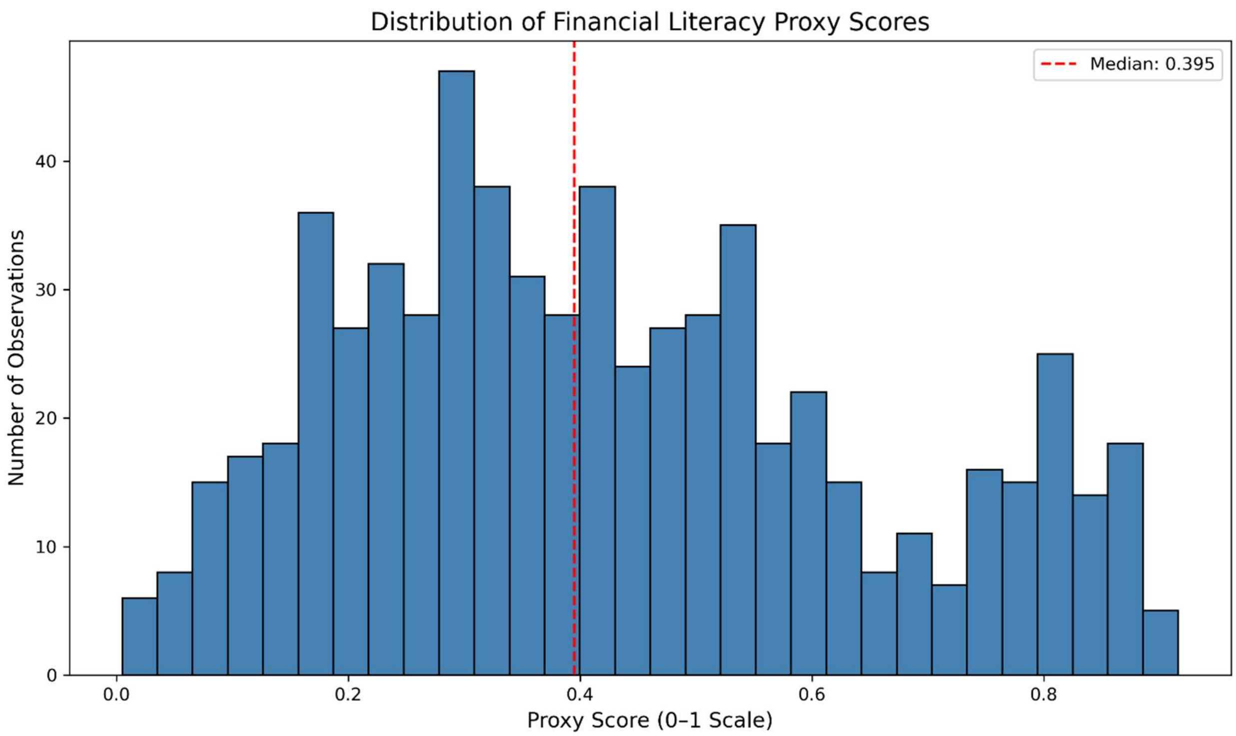

Figure 1: Distribution of Financial Literacy Proxy Scores
The proxy scores are roughly bell-shaped with a median around 0.53. This informed the binary classification threshold.

Model Evaluation

Three models were tested: **Logistic Regression**, **Random Forest**, and **XGBoost**, each evaluated with stratified 5-fold cross-validation. Accuracy, ROC-AUC, PR curves, and SHAP were used to assess performance and interpretability.

Logistic Regression

- Cross-val accuracy: 94.7% (± 2.3%)
- Interpretable coefficients show direct positive/negative influence

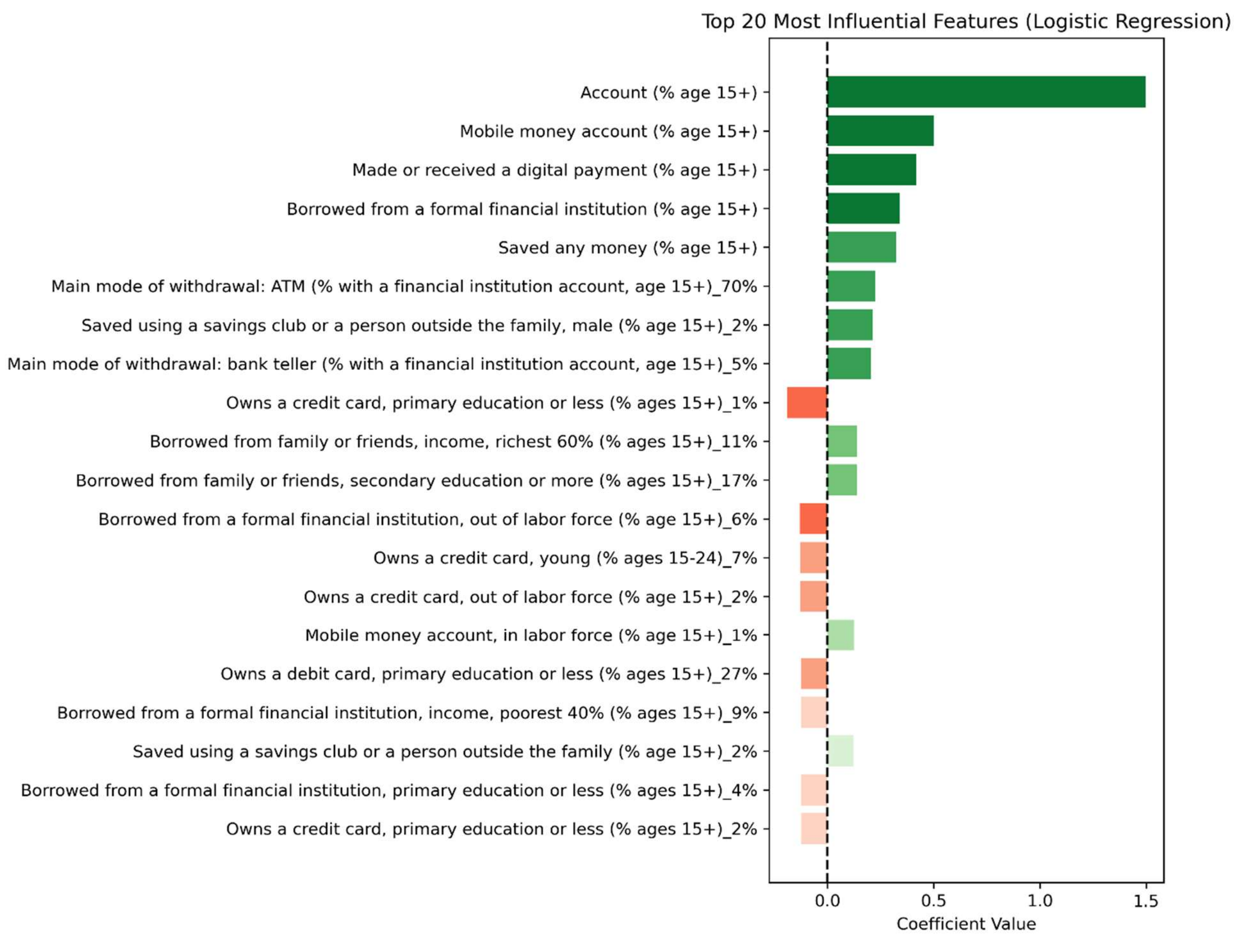

Figure 2: Top 20 Features (Logistic Regression)
Green bars indicate features positively associated with high financial literacy; red bars indicate negative association.

Random Forest

- Cross-val accuracy: 86.3% (± 3.6%)
- Captures non-linear relationships and interactions, but slightly less performant

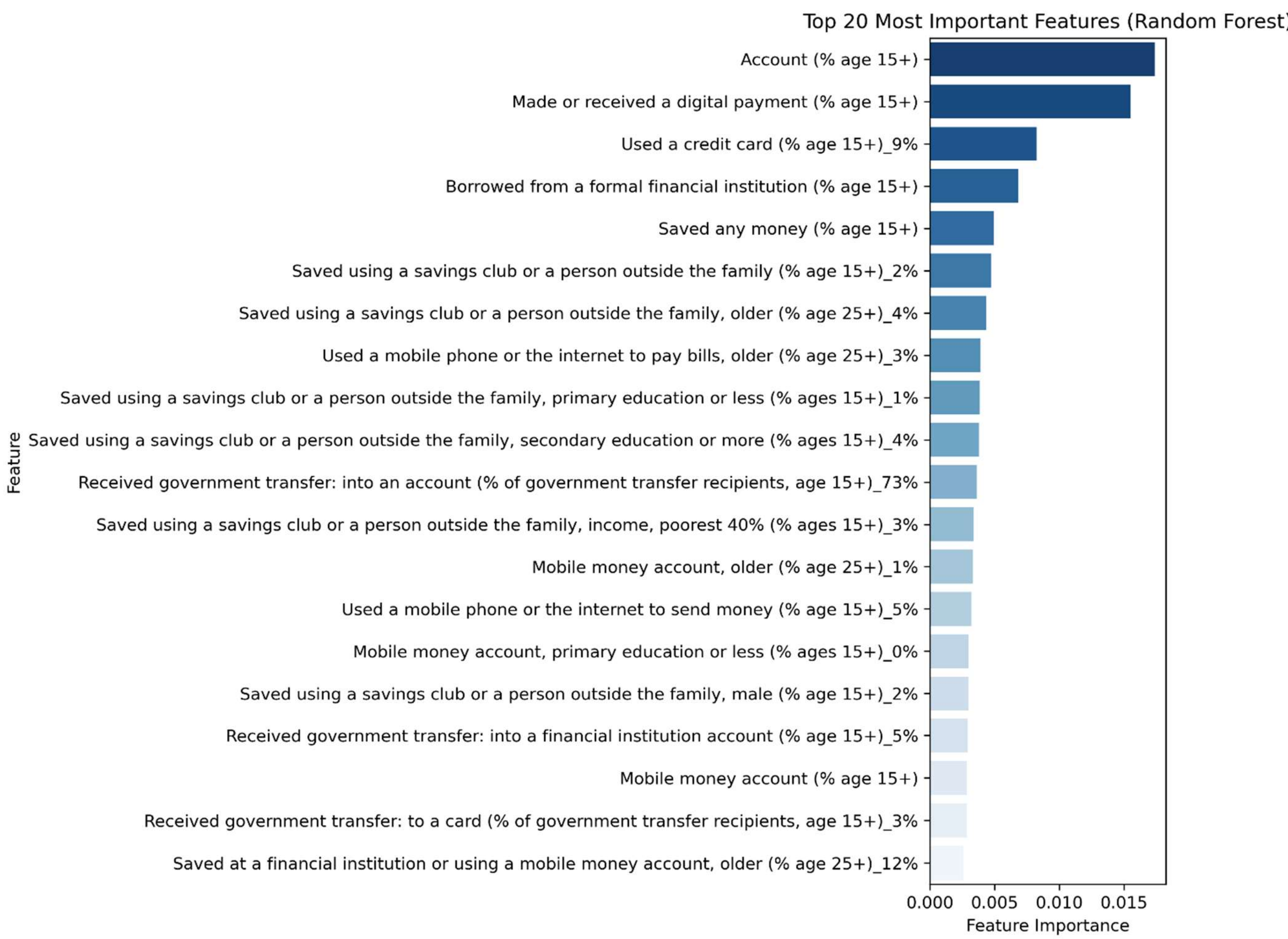

Figure 3: Top 20 Features (Random Forest)
Feature importances highlight the strongest predictors in ensemble decision paths.

XGBoost

- Cross-val accuracy: 96.0%
- Best performance overall
- SHAP explains predictions at both global and local level

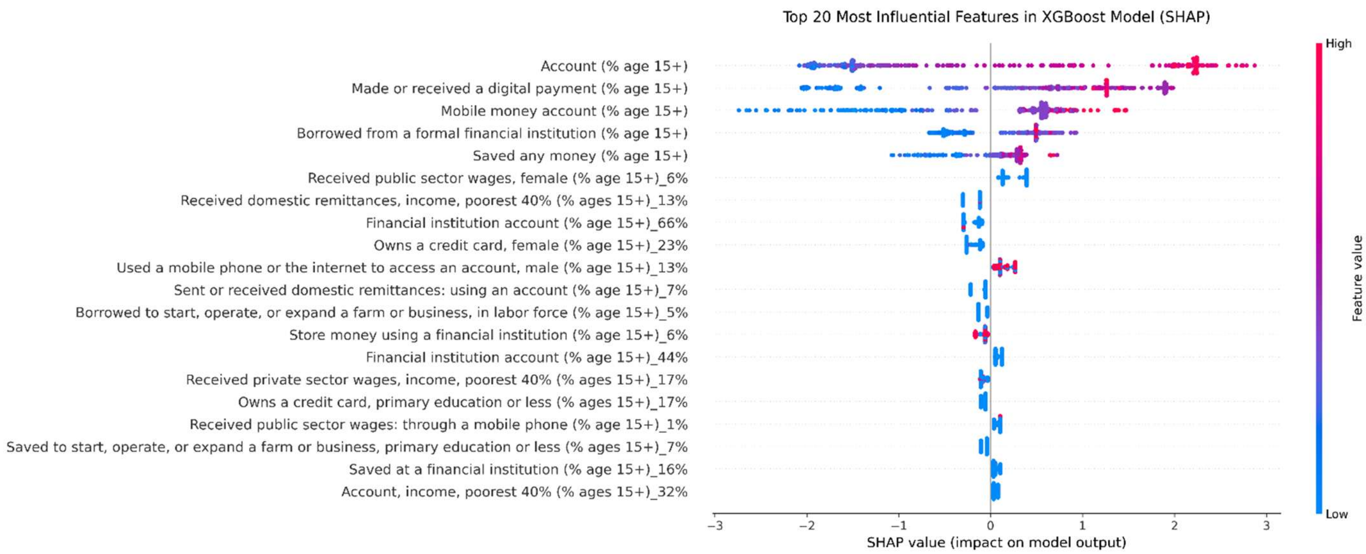

Figure 4: SHAP Summary Plot (XGBoost)
SHAP values identify key contributors across all predictions. Higher values increase predicted probability of "high" financial literacy.

SHAP Dependence Plots

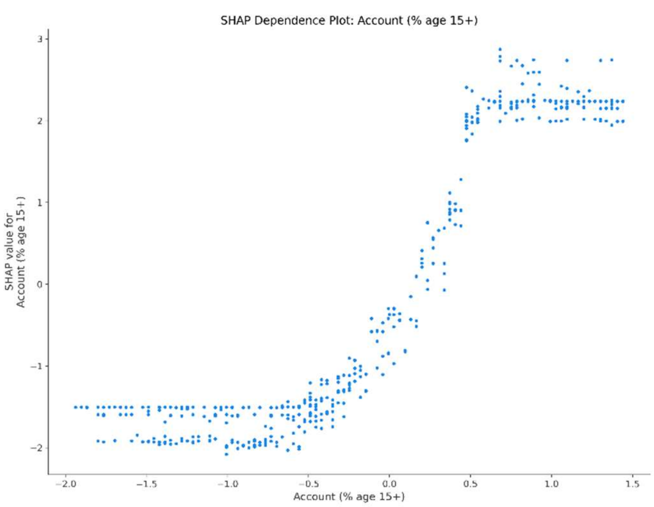

Figure 5: Account Ownership (Age 15+)

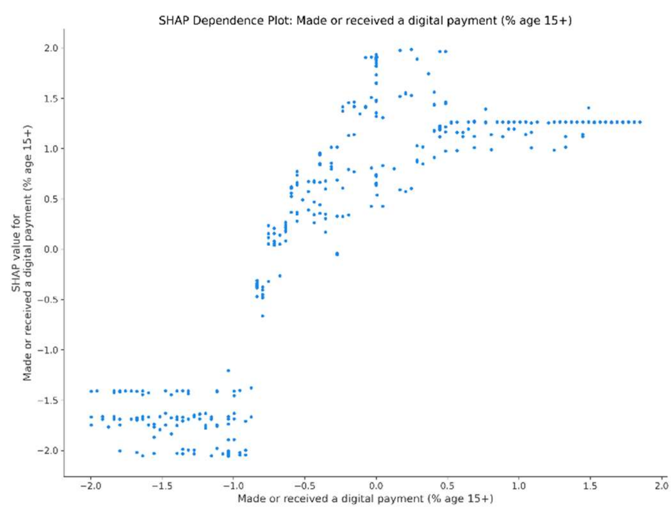

Figure 6: Digital Payments Usage

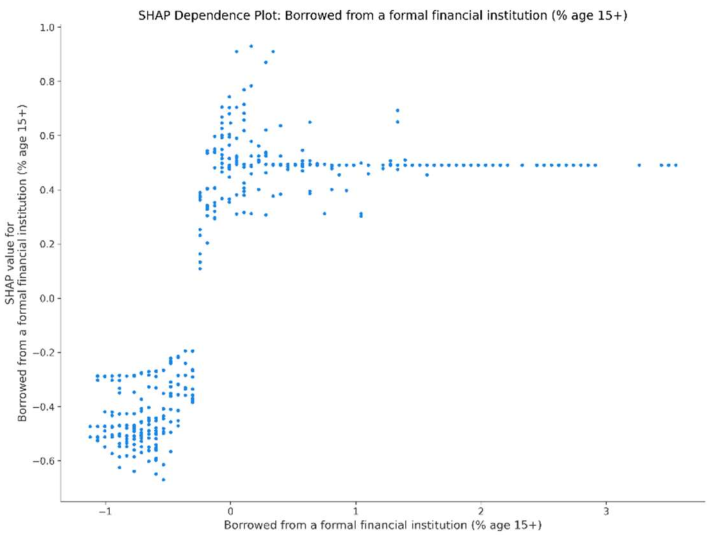

Figure 7: Savings Activity

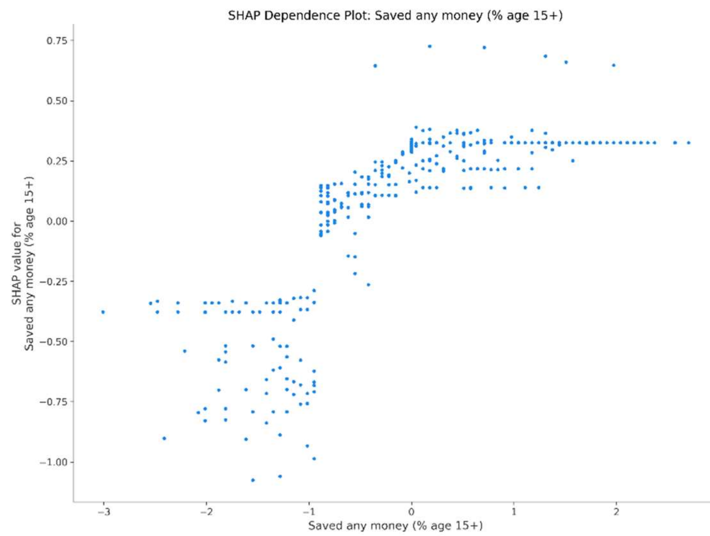

Figure 8: Emergency Fund Access

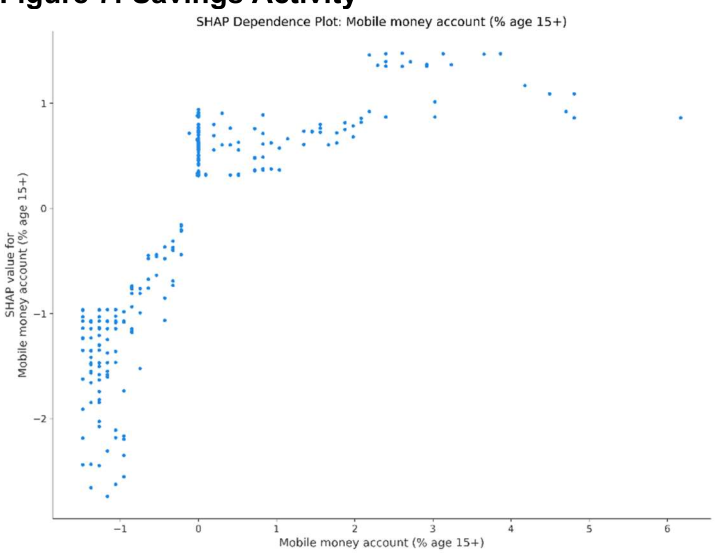

Figure 9: Mobile Money (Labour Force)

These plots show how each feature's value relates to its impact on the model's prediction. The results are consistent with earlier research highlighting the role of formal account use and mobile money services in promoting financial inclusion and literacy (Grohmann et al., 2018; Demirgüç-Kunt et al., 2018).

Business/Application Insights

These models can help identify regions or subpopulations most in need of targeted financial education. Financial institutions, including credit unions, could use this approach to improve outreach, prioritize resources, and design tailored services. The SHAP plots in particular allow practitioners to explain why certain populations are flagged as vulnerable. Yue and Zhu (2025) emphasize the potential of machine learning to cost-effectively identify low-literacy individuals, while Lu et al. (2024) highlight the utility of financial literacy indicators in improving credit risk models in agriculture, which supports the application of these techniques in other domains such as consumer banking.

Limitations and Next Steps

This project has several limitations. First, the financial literacy proxy variable was derived from behavioural indicators rather than direct assessments of financial knowledge. While useful, such a proxy may overlook conceptual understanding or attitudes toward financial decision-making (Huston, 2010). Additionally, the dataset consists of aggregated country-level data, which may obscure important within-country disparities—particularly those related to gender, income, or urban-rural divides (Haag & Brahm, 2025).

The XGBoost model, while achieving the best performance, is also more susceptible to overfitting, especially on smaller datasets. Future work could benefit from hyperparameter tuning using Bayesian optimization, or integrating regularization techniques to improve generalization. More importantly, the use of household- or individual-level data would enable a more nuanced analysis and support the creation of personalized educational interventions. Future extensions may also explore how these models perform across time or in transfer learning scenarios between countries.

Code Documentation

GitHub Repo: https://github.com/Risado8/CIND820FinalProject/tree/main

Includes final Jupyter Notebook, requirements.txt, README.md, and all visualizations

References

Demirgüç-Kunt, A., Klapper, L., Singer, D., Ansar, S., & Hess, J. (2018). The Global Findex Database 2017: Measuring financial inclusion and the fintech revolution. The World Bank.

Fernandes, D., Lynch, J. G., & Netemeyer, R. G. (2014). Financial literacy, financial education, and downstream financial behaviors. Management Science, 60(8), 1861–1883. https://doi.org/10.1287/mnsc.2013.1849

Grohmann, A., Kluehs, T., & Menkhoff, L. (2018). Does financial literacy improve financial inclusion? World Development, 111, 84–96. https://doi.org/10.1016/j.worlddev.2018.06.020

Guiso, L., Sapienza, P., & Zingales, L. (2015). How much can financial literacy help? In Financial literacy and the limits of financial decision-making (pp. 1–19).

Haag, C., & Brahm, F. (2025). The gender gap in economic and financial literacy: A meta-analysis. Applied Economics, 57(15), 1631–1650. https://doi.org/10.1080/00036846.2023.2271831

Huston, S. J. (2010). Measuring financial literacy. Journal of Consumer Affairs, 44(2), 296–316. https://doi.org/10.1111/j.1745-6606.2010.01170.x

Lokanan, M. E., Wang, H., & Lau, A. (2021). Predicting fraud victimization using classical machine learning models. Financial Innovation, 7, 1–20. https://doi.org/10.1186/s40854-021-00233-w

Lu, Z., Li, H., & Wu, J. (2024). Exploring the impact of financial literacy on predicting credit default among farmers: An analysis using a hybrid machine learning model. Borsa Istanbul Review, 24, 352–362. https://doi.org/10.1016/j.bir.2023.11.003

Lusardi, A., & Mitchell, O. S. (2014). The economic importance of financial literacy: Theory and evidence. Journal of Economic Literature, 52(1), 5–44. https://doi.org/10.1257/jel.52.1.5

Yue, A. Y. F., & Zhu, F. (2025). Unlocking financial literacy with machine learning: A critical step to advance personal finance research and practice. Technology in Society, 81, 102797. https://doi.org/10.1016/j.techsoc.2024.102797