risks

MDPI

Article

Optimizing Pension Participation in Kenya through Predictive Modeling: A Comparative Analysis of Tree-Based Machine Learning Algorithms and Logistic Regression Classifier

Nelson Kemboi Yego [1,2,3,*], Juma Kasozi [1,4] and Joseph Nkurunzinza [1,2]

[1] African Center of Excellence in Data Science, University of Rwanda, Kigali 4285, Rwanda
[2] School of Economics, University of Rwanda, Kigali 4285, Rwanda
[3] Department of Mathematics and Computing, Moi University, Eldoret 3900-30100, Kenya
[4] Department of Mathematics, Makerere University, Kampala 7062-10218, Uganda
* Correspondence: nelsonyego@gmail.com

Abstract: Pension plans play a vital role in the economy by impacting savings, consumption, and investment allocation. Despite declining mortality rates and increasing life expectancy, pension enrollment remains low, affecting the long-term financial stability and well-being of populations. To address this issue, this study was conducted to explore the potential of predictive modeling techniques in improving pension participation. The study utilized three tree-based machine learning algorithms and a logistic regression classifier to analyze data from a nationally representative 2019 Kenya FinAccess Household Survey. The results indicated that ensemble tree-based models, particularly the random forest model, were the most effective in predicting pension enrollment. The study identified the key factors that influenced enrollment, such as National Health Insurance Fund (NHIF) usage, monthly income, and bank usage. The findings suggest that collaboration among the NHIF, banks, and pension providers is necessary to increase pension uptake, along with increased financial education for citizens. The study provides valuable insight for promoting and optimizing pension participation.

Keywords: pension uptake; machine learning; tree-based models; random forest classifier

Citation: Kemboi Yego, Nelson, Juma Kasozi, and Joseph Nkurunzinza. 2023. Optimizing Pension Participation in Kenya through Predictive Modeling: A Comparative Analysis of Tree-Based Machine Learning Algorithms and Logistic Regression Classifier. *Risks* 11: 77. https://doi.org/10.3390/risks11040077

Academic Editor: Shengkun Xie

Received: 10 February 2023
Revised: 23 March 2023
Accepted: 27 March 2023
Published: 18 April 2023

1. Introduction

The role of pension schemes in the economy is crucial. They provide benefits to retirees and also impact the saving and consumption decisions of individuals and firms. Additionally, pension schemes channel periodic contributions into investments (Serrano and Peltonen 2020). Furthermore, pension uptake may increase the uptake of insurance products and the use of formal healthcare among the aging population. The importance of pension schemes extends beyond just the members and has a significant impact on the overall economy, particularly as life expectancy rises and mortality decreases. Hence, there is a need to stimulate pension participation (Balasuriya and Yang 2019; Riumallo-Herl and Aguila 2019).

As is the situation in most parts of Africa, mortality rates in Kenya have declined while life expectancy has increased. For instance, the Kenyan all-cause mortality rate reduced from 850.3 deaths per 100,000 in 1990 to 579.0 deaths per 100,000 in 2016. The under-five mortality rate, on the other hand, reduced from 95.4 deaths per 1000 live births in 1990 to 43.4 deaths per 1000 live births in 2016. The maternal mortality rate reduced from 315.7 deaths per 100,000 in 1990 to 257.6 deaths per 100,000 in 2016, with steeper declines observed after 2006. Furthermore, life expectancy at birth increased by 5.4 years, with higher gains in females than males in all but ten counties. Hence, generally, all measures of mortality experienced a decline (Achoki et al. 2019).

In the midst of declining fertility and morbidity rates, pension schemes, especially those targeting the aging population, tend to improve household welfare. The dimensions of welfare that have been observed to improve with pension uptake include an increase in monthly consumption expenditure, food expenditure, nonfood expenditures, and expenditures on household assets. This indicates an increase in the general standard of living of the pensioners concerned. In addition, there has been a reduction in labor supply, which means that older individuals will not have to struggle to find work in order to make ends meet, since they will have a steady income (Unnikrishnan and Imai 2020).

The need for an inclusive social protection that shields the population from the risk of financial hardship upon retirement is heightened by the increasing dependency ratio due to the fact of declining mortality and rising life expectancy. Despite its important role, pension uptake in Africa is low, with the pension coverage of the various schemes in the region extending to only a small portion of the population, mostly those involved in the formal sector, leaving a large part of the population uncovered. This situation has been partly attributed to the failure of the contributory pension system, which is widely used in the region, to respond to the needs of the majority of the population involved in the informal sector. As a result, a large portion of the population is ineligible for any pension benefits upon retirement. Moreover, the coverage gap among the elderly may persist in most countries into the foreseeable future. It has been common for the elderly to be supported by the youthful working population. However, due to the fact of rural to urban migration, which is common among the younger working population, the elderly may end up with fewer resources and face abject poverty (Guven 2019).

The situation in Kenya, which is part of the region, is not any better. The pension uptake is also biased, covering mainly high-income earners in formal employment, leaving a large portion of the population uncovered (Künzler 2016). Moreover, conventional pension schemes in the country have not been able to attract a significant number of clients. This is despite Kenya having made tremendous improvements in financial inclusion, rising from 26.7% in 2013 to 82.9% in 2019, representing one of the highest levels of financial inclusion in the region (Central Bank of Kenya et al. 2019a; Asuming et al. 2019).

Despite efforts by the government through the retirement benefits industry's regulator to increase pension uptake, it remains low. Some of the measures taken include well-designed marketing campaigns, as well as the introduction of pension programs for those in the informal sector. Other pension schemes, such as the National Social Security Fund (NSSF), are also accessible options. One of the major programs introduced by the regulator, aimed at the informal sector, was the "Mbao Pension Scheme". The program was designed to be affordable and flexible for the informal economy (Kwena and Turner 2013).

Despite the focus on the formal sector, none of the pension schemes, including those managed by private insurers, have achieved significant uptake. As a result, they cover less than 15% of the working population, leaving most Kenyans without any income in retirement. The low coverage has been attributed to a lack of participation from professional workers (Consumer Options Ltd. 2019). As a result, Kenya's social security policies have failed to provide inclusive social protection for the population through any form of social protection or retirement benefit scheme, causing financial hardship and poverty among the elderly. Therefore, it is crucial to identify the determinants of pension uptake among individuals using recent data. Doing so would facilitate the timely intervention towards ensuring optimum pension participation.

The phenomenon of low adoption rates for pension and long-term care insurance products is not exclusive to a single country but is widespread across various regions worldwide. Some developed markets have also been experiencing this challenge (Hadad et al. 2022). The UK, for instance, has seen a decline in pension participation in recent times (Balasuriya and Yang 2019). However, the negative impact of low uptake is particularly pronounced in developing economies due to the fact of their resource-constrained markets, which often lack the necessary infrastructure, ecosystem, policies, and resources to facilitate long-term financial planning. Consequently, citizens in these economies are more suscepti-

ble to financial insecurity during old age or times of illness. This susceptibility is further exacerbated by ongoing global crises, such as climate change and the COVID-19 pandemic. Additionally, cultural factors and inadequate awareness of the benefits of long-term care insurance products may also contribute to the low uptake of such financial products (Rajan et al. 2023; Pörtner et al. 2022; Guerrero et al. 2021).

Previous analyses of pension participation have employed probit regression models (Balasuriya and Yang 2019; Lades et al. 2017). The current study adopted three tree-based models and a logistic regression classifier and compared their performance before selecting the optimal model for the final analysis. Machine learning models are better suited to identify nonlinear relationships between variables, which is particularly useful in the study of pension participation. There are many factors that can influence an individual's decision to enroll in a pension plan, and traditional statistical models may be limited by their assumptions and modeling techniques. Machine learning algorithms can analyze these variables and their interactions in a more comprehensive and flexible way, and they can learn and adapt from new data over time, making them valuable in the study of pension participation.

Pension participation rates have been low, which leaves much of the populace vulnerable to financial insecurity during old age or times of illness. While previous studies have employed traditional statistical models to understand the factors that influence pension participation, these models may not be able to capture the complex nonlinear relationships between the various factors. Therefore, there is a need for a more comprehensive and flexible approach that can better identify the factors that influence pension participation and predict an individual's likelihood of enrolling in a pension plan.

To address the problem of low pension participation rates, a predictive model is needed to identify the key factors that influence an individual's decision to enroll in a pension plan. This predictive model should be able to analyze a wide range of demographic and economic variables and their interactions to accurately predict an individual's likelihood of participating in a pension plan. By accurately identifying the factors that influence pension participation, policymakers and financial institutions can design targeted interventions and strategies to increase pension participation rates and improve financial security.

This paper makes significant contributions to academia and practice in several areas. Firstly, it demonstrates the potential of predictive modeling techniques, specifically ensemble tree-based models, in improving pension participation. It compared four machine learning models and identified the most robust for predicting pension participation. Secondly, this study identified key factors that influence pension enrollment, which is important for policymakers and other stakeholders when designing interventions and strategies to increase pension uptake. Thirdly, this study suggests strategies to increase pension uptake. Overall, the study provides valuable insights for promoting and optimizing pension participation in a resource-constrained environment (such as the case in Kenya), and the findings are likely to be relevant to other countries facing similar challenges.

This paper is organized in such a way that there are two sections at the end of this introduction which review the related literature. Section 2.1 discusses the determinants of pension uptake, while Section 2.2 explores the applications of machine learning in related areas. The paper then proceeds to Section 3, which outlines the materials and methods utilized in the study. This section provides an explanation of the data source, feature selection, and modeling techniques used in the study. The next section, Section 4, presents the results and discussions of the study, which is divided into several subsections evaluating the performance of different models and feature importance. Finally, the paper concludes with a summary of the main findings and their implications. Appendix A is also included, which provides additional details.

2. Related Literature

2.1. Determinants of Pension Uptake

Some of the factors that have been previously found to influence the uptake of pension or other forms of old-age social protection include income, education level, work-related associations, and age (Guven 2019; Kitheka 2020). Moreover, gender, place of residence, and occupation have also been found to influence pension uptake (Kibona 2020). Riumallo-Herl and Aguila (2019) posit that pension uptake could encourage the uptake of other social programs, such as health insurance, but they do not rule out the interaction effect between social health schemes and pension schemes (Riumallo-Herl and Aguila 2019). As an additional factor, personal traits have been noted to contribute to life outcomes. In particular, self-control has been found to be associated with factors such as home ownership, education, and economic status. Therefore, the higher the level of self-control during childhood, the higher the level of pension participation (Cobb-Clark et al. 2022; Lades et al. 2017).

It has been found that there is a correlation between the level of income and the uptake of pensions. As an individual's income increases, their willingness to participate in pension schemes also increases. Similarly, pension uptake has been found to correlate with an increase in education level. An increase in education level is said to correlate with financial literacy, which in turn correlates with a willingness to take up some form of retirement social protection scheme. This contributes to an increase in the saving and investment culture, as well as a rise in sound financial management and the level of financial culture. Hence, both the level of income and the highest level of education attained have a positive correlation with pension uptake (Guven 2019; Kitheka 2020).

In relation to age, a positive association has been observed between age and the uptake of pensions. For example, individuals above the age of 35 have been found to save more compared to those below the age of 35. However, it has been found that the younger urban population is more willing to take up micro-pension plans, indicating that the interaction of ability to pay cannot be disregarded. Nevertheless, assuming all other factors remain constant, a positive correlation between pension uptake and age has been observed (Kitheka 2020).

Workplace association has also been found to positively correlate with the uptake of pensions. On the basis of gender, males have been observed to have higher levels of pension uptake compared to females. On location, urban dwellers have been said to have higher levels of pension uptake compared to their rural counterparts. On occupation, individuals involved in either farming or fishing have been said to be less likely to take up pension covers compared to those involved in formal employment (Kibona 2020; Kitheka 2020).

2.2. Applications of Machine Learning in Related Areas

The application of machine learning techniques in finance has been gaining popularity due to the growth of data and advancements in the field (Levantesi and Zacchia 2021; Kipkogei et al. 2021; Dixon et al. 2020). Machine learning has been found to provide insights that are not attainable through traditional parametric methods and is more flexible for handling high dimensional data. The rise of fintech as an industry has been attributed to the development of machine learning, the growth of data, the increased mobile usage, the rise in digital payments, APIs, and the increase in the amount of capital available (Dixon et al. 2020).

Machine learning has been used in various financial applications, such as behavioral prediction, price modeling, algorithmic trading, portfolio management, fraud detection, customer churn, investor sentiment analysis, and credit risk prediction (Dixon et al. 2020; Renault 2020; Belhadi et al. 2021). Bankruptcy prediction, cryptocurrency volatility prediction, clustering causes of death in insurance-related data (Bett et al. 2022), predict farmers' uptake of crop insurance (Mare et al. 2022), and business sustainability have also been modeled using machine learning techniques (Bouri et al. 2021a; Barboza et al. 2017; Kipkogei et al. 2021).

In this study, four traditional machine learning classifiers were used to find a model that robustly predicts the uptake of pensions. Tree-based models have been found to perform optimally for tabular financial data (Bouri et al. 2021a; Kipkogei et al. 2021; Levantesi and Zacchia 2021). Although deep learning classifiers have demonstrated high performance in high-dimensional spaces, simpler classifiers have been shown to outperform them for tabular data (Yego et al. 2021; Renault 2020). Machine learning has been applied in the prediction of early retirement in the retirement benefits sector (Boado Penas et al. 2019). However, there has been no attempt to use it for predicting pension uptake.

3. Materials and Methods

This study utilized machine learning classification models on nationally representative data to predict pension uptake. It compared four classification models and selected the most robust for the prediction task. The data were used to train the models and evaluate their performance in predicting pension uptake.

3.1. Data

This study utilized nationally representative data from the 2019 Kenya FinAccess Household Survey. The survey aimed to provide research data and measure access to and demand for financial services. The survey used a nationally representative cross-sectional design with a multistage, stratified cluster sampling method. A total of 8669 interviews were conducted across 820 clusters, with one person interviewed per household. The data collected included information on sociodemographic characteristics, access and usage of financial services, as well as mobile money and pension uptake (Central Bank of Kenya et al. 2019b). A broader description of the data may be found in Yego et al. (2021).

3.2. Feature Selection

It is necessary to provide a feature selection criterion that can assess how important each feature is to the output class and labels (Chandrashekar and Sahin 2014). The initial data had more than a thousand variables but only 30 sociodemographic variables. Those rejected were either not sociodemographic variables, irrelevant, or were redundant in that they were explained fully or to a great extent by one or more of the accepted variables.

3.3. Data Split

The data were split such that the training data were not used in either model validation or testing. The train–test split had a ratio of 70:30, with the majority of the data (70%) being used for model training and the remaining 30% used for validation and testing. The test–validation split, on the other hand, was 1:1, resulting in a final train–validation–test ratio of 0.7:0.15:0.15. This means that 15% of the data were used for hyperparameter tuning and were not used to test the model's ability to perform with "unforeseen" data.

Most scholars agree that a significant portion of the data should be used for training. Levantesi and Zacchia (2021) and Mutai et al. (2021) used a 70%:30% train–test split. The cross-validation method used in the current study was k-fold cross-validation, with k equal to 5.

3.4. Hyperparameter Optimization

Hyperparameter optimization was performed to identify the parameters that would result in the optimal performance for each model. Tables A3 and A4 display the hyperparameters that were tuned for the random forest classifier and XGBoost classifier, respectively. For the random forest model, the optimized parameters were n_estimators (number of trees in the forest) at 110, max_features at "auto", min_samples_split at 2, and bootstrap set to true. For the XGBoost model, the optimized parameters were n_estimators at 1000, max_depth set to "auto", max_features at 0.9, and gamma at 0.1.

3.5. Model Training

Figure 1 illustrates the process followed in the training and testing of the model. The aim was to find an optimal classifier, h, that predicts the likelihood of a potential client taking up a pension plan based on a set of label Y (binary: pension uptake or nonuptake) and a sequence domain set of features X. The loss Ls(h) in the test set was minimized such that h: $X \rightarrow Y$, where H is the hypothesis class expected to contain h (Shalev-Shwartz and Ben-David 2014).

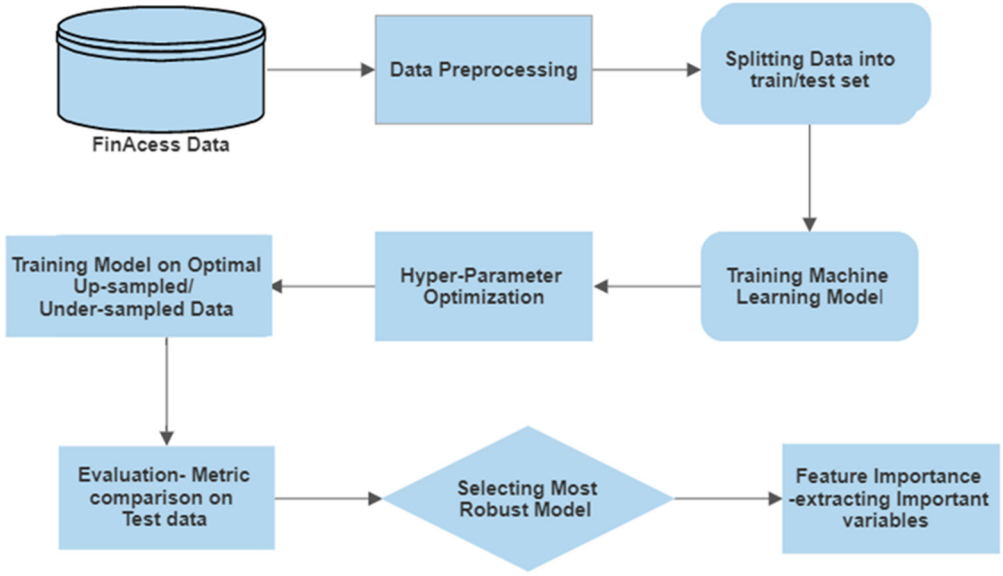

Figure 1. General machine learning process followed.

3.6. Models Trained and Tested

Four machine learning classifiers were considered for predicting pension uptake: decision tree, random forest, extreme gradient boosting (XGBoost), and logistic regression. In the context of the current study, outcome *Y* refers to the predicted variable being investigated, which is whether the pension was taken up by the participants. This variable is expressed in binary form, meaning it can only take one of two possible values: one or zero. The value of one indicates that the participant has taken up the pension, while the value of zero indicates that they have not.

$$Y = \begin{cases} 1 & = \quad Pension\ uptake \\ 0 & = \quad Pension\ non-uptake \end{cases} \tag{1}$$

This binary representation of the outcome variable is often used in statistical analysis to simplify the data and facilitate computation. By using a binary variable to represent the uptake of the pension, researchers can more easily compare and analyze the factors that contribute to or influence this outcome (Shalev-Shwartz and Ben-David 2014).

3.6.1. Logistic Regression Classifier

Logistic regression is a widely used classification and regression method for binary prediction outcomes. It has shown high accuracy in several studies, including cancer survival prediction (Kutrani et al. 2021) and the prediction of drivers of preterm birth (Saroj and Anand 2021). In the latter study, the logistic regression classifier outperformed the decision tree classifier, exhibiting a higher precision score, f1 score, and AUC. This underscores the potential of logistic regression in improving prediction accuracy and enhancing decision making in various applications.

Logistic regression was part of the generalized linear models, with the response variable generalized using a logit link following a binomial distribution with binary outcomes.

Logistic regression has been presented as part of a wider class of generalized linear models. However, unlike the linear regression model, the response variable is generalized with a link, which was a logit link for this study, following a binomial distribution with a binary outcome.

For the set features $\boldsymbol{X} = x_1 \ldots x_n$, the probability of pension uptake is given by:

$$Probability = E(Y|x_1 \ldots x_n) = \frac{exp\{\beta_0 + \beta_1 x_1 + \ldots + \beta_n x_n\}}{1 + exp\{\beta 0 + \beta 1 x \ldots\ldots + \beta_n x_n\}} \quad (2)$$

where $\beta_1, \ldots\ldots \beta_n$ are the respective estimated coefficients, and β_0 is the intercept (Dixon et al. 2020; Levantesi and Zacchia 2021).

3.6.2. Decision Tree Classifier

Among the four models compared, three models (decision tree, random forest, and XGBoost) are tree-based models. Decision trees predict the label Y associated with instances X by navigating from a root node to a leaf node. At each node, the successor child is chosen based on the splitting of the input space, with the best attribute chosen based on some criterion, such as entropy or Gini. The attribute is used to create nodes and divide the data into subsets, and the process is repeated until class purity is achieved. Although decision trees are easy to interpret, they are more likely to attain a local optima than ensemble methods (Levantesi and Zacchia 2021).

3.6.3. Random Forest and Extreme Gradient Boosting (XGBoost)

Random forest is an ensemble model that averages over a collection of decision trees, while XGBoost is an ensemble model that boosts decision trees (Dixon et al. 2020). The trees in a Random Forest model are drawn from the same distribution, although they are sampled identically and independently (Breiman 2001). Random forest has been shown to exhibit better accuracy and robustness in predicting insurance fraud, making it a plausible model for insurance uptake prediction (Li et al. 2016). The random forest model creates multiple classification trees, and these trees' predictions are averaged to estimate the classification function. This produces a combined output denoted as:

$$f_{av}(X) = \frac{1}{N} \sum_{n=1}^{N} f_n(X) \quad (3)$$

where f_n is the prediction obtained from training a classification tree on the nth new dataset (Diana et al. 2019).

The XGBoost algorithm operates by starting with identical initial predictions. Next, it constructs a decision tree by analyzing the pseudo-residuals of each sample and selecting the partition with the highest gain. This gain is calculated by adding up the similarity scores of the left and right child nodes and subtracting the similarity score of the parent node. The trees are then pruned, and each node's output value is computed in terms of log-odds. The predicted probabilities are updated by iterating the process and converting the log-odds to probabilities again (Bentéjac et al. 2021).

3.7. Handling Class Imbalance

Class imbalance arises when the training data in one class are proportionally larger than another class. The larger class is referred to as the majority class, while the smaller class is the minority class. When a machine learning model is trained on imbalanced data, the trained model tends to capture the bias inherent in the classes; hence, the model performance metrics would be lower than if the classes were balanced before training.

As shown in Figure 2, there was a significantly higher number in the non-pension uptake class compared to the pension uptake class. The pension uptake class had 1107 observations (representing 12.77% of the total) compared to the non-pension uptake class which had 7562 (representing 87.23% of the total). Therefore, the data used were imbalanced,

and there was a need to handle the class imbalance problem. To handle this, resampling was used in this study. This was performed by either adding copies of instances from the under-represented class (i.e., up-sampling) or by deleting instances from the over-represented class (i.e., down-sampling). Both methods were performed simultaneously to enhance the robustness of the results. Figure 3 shows the class balance after the data were balanced through up-sampling. Similarly, the class balance after down-sampling is also shown but with fewer instances for the down-sampled data, as instances of the majority class were randomly deleted to balance with the minority class.

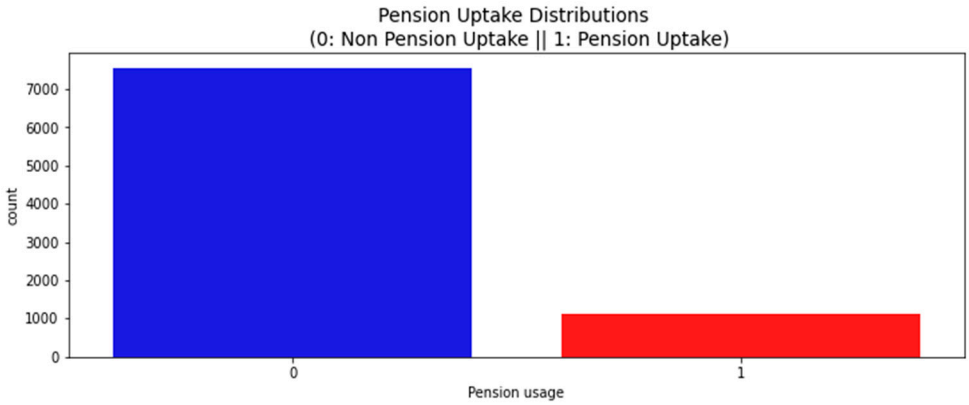

Figure 2. Data Imbalance.

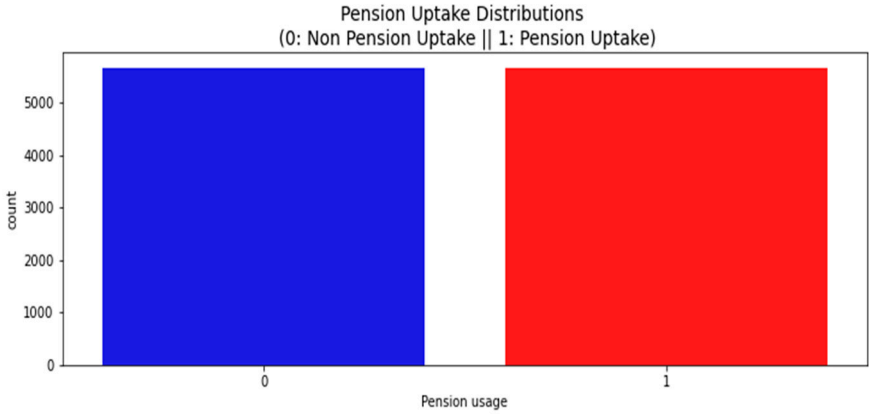

Figure 3. Class balance after up-sampling of the data.[1]

4. Results and Discussion

4.1. Descriptive Data Analysis

As shown in Table A2, there were 30 variables, each with 8669 observations. None of the variables had missing values or duplicate rows. Among the 30 variables, 15 were categorical, 11 were Boolean with binary answers, and the remaining 4 were numeric. The results presented are from the test dataset. After preprocessing and feature selection, 22 features were retained for the final training, excluding the dependent variable.

Figures A4 and A5 display the connection between different variables and pension uptake using the dataset that majorly comprised categorical variables and few continuous variables. To examine the associations between these variables, a correlogram was generated using the Cramer's V correlation coefficient, which measures the association between categorical variables (Barrera Ferro et al. 2020). A Cramer's V correlation coefficient function was utilized to generate a correlation matrix for the categorical columns, and Seaborn's heatmap function was employed for correlogram visualization. The results in Figure A4 indicate that NHIF utilization had the most significant correlation with pension adoption, with bank usage and educational background following closely behind. This finding is in line with the intuitive notion that banks provide a desirable and reliable option

for retirement savings accounts. Moreover, individuals with higher levels of education are more cognizant of the importance of pension savings and possess the means to make contributions to such plans.

4.2. Evaluation Metrics on Unbalanced Data

Table 1 shows the evaluation metrics (precision score, recall score, F1 score, and accuracy) for various models based on the raw unbalanced data. The accuracy of the ensemble tree-based models was higher, with random forest having an accuracy of 0.904762 and XGBoost having an accuracy of 0.909370, compared to the logistic regression classifier, which had an accuracy of 0.878648, and the decision tree classifier, which had the lowest accuracy of 0.857143. Therefore, the ensemble tree-based classifiers are expected to provide a higher ratio of correctly classified observations compared to standalone models for this type of data. This implies that despite the data imbalance, the ensemble tree-based models (random forest and XGBoost) provided a relatively higher level of accuracy compared to standalone classifiers.

Table 1. Evaluation metrics for unbalanced data.

Index	Model	Precision Score	Recall Score	F1 Score	Accuracy
0	Logistic Regression	0.749632	0.604775	0.634678	0.878648
1	Decision Tree	0.684142	0.671249	0.677285	0.857143
2	Random Forest	0.838286	0.693754	0.738656	0.904762
3	XGBoost	0.819716	0.750628	0.779006	0.909370

XGBoost had the highest F1 and recall scores (0.779006 and 0.750628, respectively), followed by random forest (0.738656 and 0.693754), decision tree classifier (0.677285 and 0.671249), and the logistic regression classifier (0.634678 and 0.604775). In terms of precision, random forest had the highest score (0.838286), followed by XGBoost (0.819716), logistic regression classifier (0.749632), and decision tree classifier (0.684142). This implies that the random forest classifier is the most robust in providing the positive class.

4.3. Evaluation Metrics on Up-Sampled Data

Table 2 shows the evaluation metrics (precision score, recall score, F1 score, and accuracy) for the various models' balanced up-sampled data. Based on the table, the random forest model has the highest precision score of 0.98058, recall score of 0.97786, F1 score of 0.97878, and accuracy of 0.97887 compared to the other models. This suggests that the random forest model outperforms the other models in accurately predicting the positive class and avoiding false positives and negatives. All of the models, except logistic regression, improved their accuracy when the data were up-sampled compared to the raw unbalanced data. The ensemble tree-based classifiers had higher accuracy than their standalone counterparts. This is consistent with previous findings (Kipkogei et al. 2021; Yego et al. 2021), where ensemble methods showed better performance than standalone models.

Table 2. Evaluation metrics on up-sampled data.

Index	Model	Precision Score	Recall Score	F1 Score	Accuracy
0	Logistic Regression	0.80274	0.80337	0.80270	0.80282
1	Decision Tree	0.97293	0.96864	0.96990	0.97007
2	Random Forest	0.98058	0.97786	0.97878	0.97887
3	XGBoost	0.96194	0.95588	0.95744	0.95775

Table 2 further shows that random forest had the highest F1, recall, and precision scores (0.957443, 0.955881, and 0.961942 respectively), followed by XGBoost (0.955656, 0.954036, and 0.960484), decision tree classifier (0.941310, 0.939276, and 0.949136), and

logistic regression (0.785855, 0.783468, and 0.784030). Based on the F1, recall, and precision scores, random forest was the most robust model when the data were up-sampled, and the tree-based ensemble models showed higher scores than the nonensemble algorithms.

4.4. Evaluation Metrics on Down-Sampled Data

Table 3 displays the evaluation metrics (precision score, recall score, F1 score, and accuracy) for the models trained on down-sampled data. XGBoost had the highest performance in all metrics (0.957746, 0.957443, 0.955881, and 0.961942 respectively), followed by the random forest classifier (0.952465, 0.952078, 0.950346, and 0.957595), then the decision tree classifier (0.945423, 0.944906, 0.942966, and 0.951922), and finally, the logistic regression classifier (0.852113, 0.851848, 0.851964, and 0.851754). Similar to the up-sampled data, the ensemble machine learning models showed higher scores than the standalone algorithms. However, XGBoost outperformed the random forest classifier by a slight margin in all metrics.

Table 3. Evaluation metrics on down-sampled data.

Index	Model	Precision Score	Recall Score	F1 Score	Accuracy
0	Logistic Regression	0.851754	0.851964	0.851848	0.852113
1	Decision Tree	0.951922	0.942966	0.944906	0.945423
2	Random Forest	0.957595	0.950346	0.952078	0.952465
3	XGBoost	0.961942	0.955881	0.957443	0.957746

4.5. Areas under Receiver Operating Characteristic Curves

Figure 4 displays the AUCs (area under the curve) of the logistic regression, decision tree, random forest, and XGBoost classifiers for the up-sampled data. The dotted, diagonal line in the figure represents the 0.5 mark, which is the point at which the AUC would be equivalent to a random guess or a fair coin toss. All areas represented by the four models are better than a random guess, with random forest having a very high AUC of 0.9999 in distinguishing between pension uptake and nonuptake. The second highest AUC is shown by XGBoost, with a value of 0.9925, followed by decision tree (0.9797) and logistic regression (0.9111).

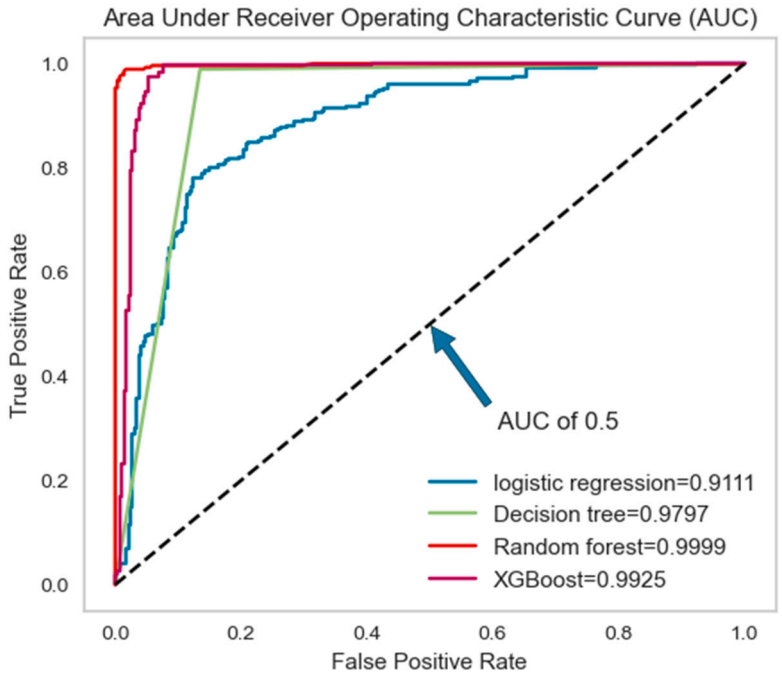

Figure 4. AUCs for up-sampled data.

The AUCs obtained from the models trained on the up-sampled data reveal that random forest is the most robust model in differentiating between pension uptake and nonuptake, as it has the highest area under the receiver operating characteristic curve. This finding differs from studies by Boado Penas et al. (2019) and Levantesi and Zacchia (2021), who found that logistic regression had a higher AUC than random forest. However, it agrees with the results reported by Kipkogei et al. (2021) and Yego et al. (2021), who found that tree-based algorithms outperformed the logistic regression classifier in terms of metric performance. Figure 5 shows the AUCs for the logistic regression, decision tree, random forest, and XGBoost classifiers for the down-sampled data. Like the up-sampled data, the AUCs for the four considered models were at least better than a random fair coin toss. However, the AUCs were lower compared to those of the up-sampled data. Random forest had an AUC of 0.9460, followed by XGBoost (0.9288), then logistic regression (0.9283), and finally, decision tree (0.7763).

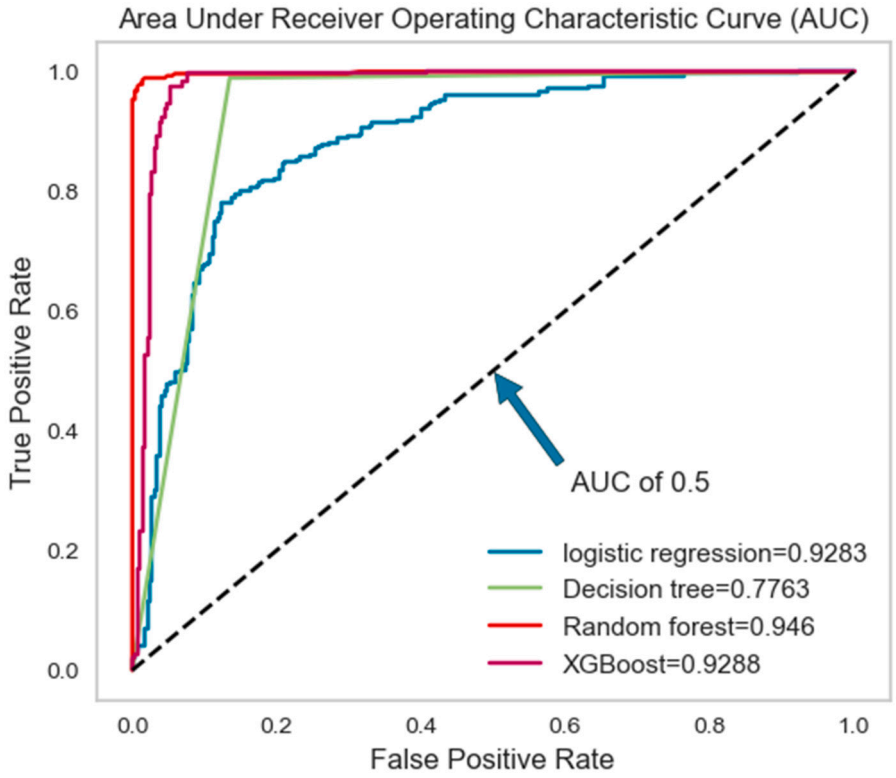

Figure 5. Areas under receiver operating characteristic curves for down-sampled data.

Based on the AUCs in Figures 4 and 5, the random forest classifier is the most robust in predicting pension uptake, since it had the highest chance of differentiating between pension uptake and nonuptake. The ensemble methods show better AUCs compared to the standalone models. The AUC for random forest was, furthermore, consistently the highest for both up-sampled and down-sampled data. Ensemble methods, therefore, tend to have better "collective" intelligence from either bagging or boosting (Belhadi et al. 2021). The results of our study support the findings of Lin et al. (2017) in that random forest is a reliable model for analyzing insurance big data compared to support vector machines and other classification algorithms. Similarly, Li et al. (2016) also reported that random forest had better accuracy and robustness in predicting insurance fraud.

4.6. Trends in the Pension Uptake

Figures A1–A3 display the levels of pension participation over time. Figure A2 displays the levels that have been disaggregated by gender, while Figure A3 shows the disaggregation by residence. Figures A1–A3 highlight some trends in pension participation

over time, as well as differences in participation between different residence groups. One notable finding is that overall participation in pensions has fluctuated over time, with an increase from 2006 to 2016, a decline between 2016 and 2019, and a subsequent resurgence. However, the level of participation has remained relatively low and has never exceeded 12.5%. This suggests that there may be some challenges or barriers to pension participation that need to be addressed in order to encourage more people to save for retirement. When disaggregated by gender, the data reveal that men have consistently had higher levels of pension participation than women. This could be due to the presence of a variety of factors, such as differences in access to formal employment or differences in income levels. Encouraging more women to participate in pension schemes could be an important step towards reducing gender inequality in retirement income and addressing the gender pension gap.

Figure A3 highlights disparities in pension participation between urban and rural areas. As previously found by Kitheka (2020) and Kibona (2020), pension participation has consistently been higher in urban areas. For instance, in 2019, the pension uptake was 19.6% in urban areas and 6.6% in rural areas. This suggests that factors such as infrastructure, access to financial services, or employment opportunities may be influencing participation levels. Addressing these disparities and finding ways to encourage more rural residents to save for retirement could help to improve overall pension coverage across the country. Overall, the data presented in these figures highlight some important trends and differences in pension participation in Kenya. Addressing the challenges and barriers that are preventing more people from saving for retirement, particularly women and rural residents, could have important implications for promoting financial security in old age and reducing inequality.

4.7. Feature Importance

Figures 6–9 show feature importance extracted from the random forest, XGBoost, decision tree, and logistic regression classifiers. The figures, with the exception of the variable importance extracted from the logistic regression classifier, seem to show little deviation in the importance they place on the features. For instance, all of the figures put NHIF usage as the most important feature. Therefore for the discussion, this paper adopted the order on the most robust model: random forest classifier. Table A1 and Figure 6 show the feature importance that was extracted from the random forest classifier. The features according to their importance were NHIF usage, monthly income, bank usage, poverty vulnerability, supporting others, education level, household size, age group, access to the Internet, source of financial advice, financial health, wealth quintile, ownership of the place of residence, marital status, savings in the previous 12 months, gender, whether one met one's own goals, nature of residence—whether rural or urban, vulnerability index, keeping money aside for specific purpose, social network device used, mobile ownership, and cryptocurrency usage.

Table A1 and Figure 6 demonstrate that those who participated in the NHIF program were more likely to also enroll in a pension scheme. This highlights that NHIF participation is a significant factor in pension uptake. However, as seen in Riumallo-Herl and Aguila (2019), there may be a relationship between NHIF usage as social health insurance and pension uptake. This study supports the idea that promoting pension schemes could aid in achieving universal health coverage. Therefore, a combined approach with complementary NHIF policies and pension schemes may increase enrollment in both.

Likewise, the higher the monthly income, the more likely it is for one to be enrolled in a pension scheme.

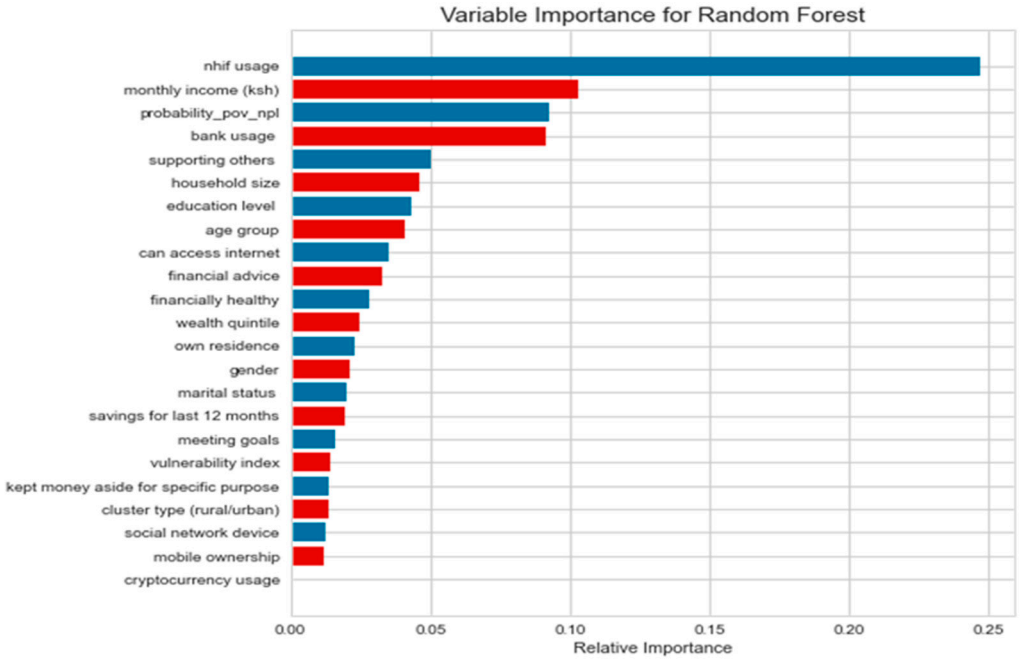

Figure 6. Feature importance from random forest.

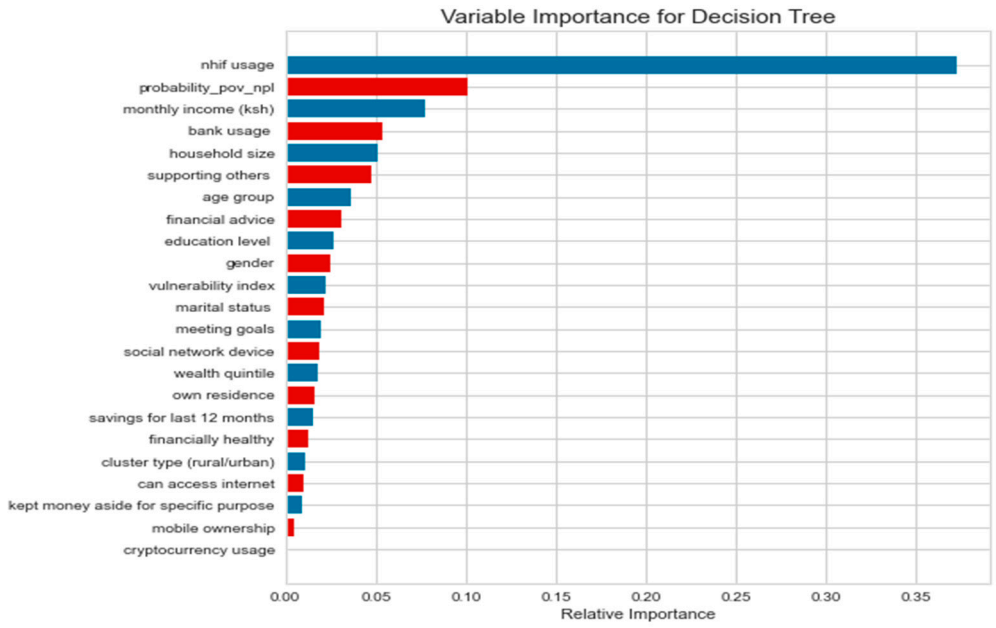

Figure 7. Feature importance from decision tree.

Moreover, the banked part of the populace had higher uptake than the unbanked populace. Similarly, those who supported others financially had a higher pension uptake than those who did not. Level of income, being banked, and support for others are pointers to financial capacity. This could be interpreted from an ability perspective, in that those who had sufficient resources as to spare some for others possibly had some to put aside in the form of a pension. This concurs with Kitheka (2020) and Guven (2019), who found pension uptake to correlate with the level of income such that as an individual's income increases, the willingness and ability to register and contribute to a pension scheme increases. Pension uptake also rose with the education level. Those who had a university-level education had higher uptake compared to those who had primary level as the highest level of education attained. This seems to concur with Kitheka (2020) and Guven (2019), who found that the

pension uptake rises with the rise in the level of education among individuals. This could imply that as the level of education rises, so does the financial literacy, which in turn instills a saving and investment culture, as well as sound financial management.

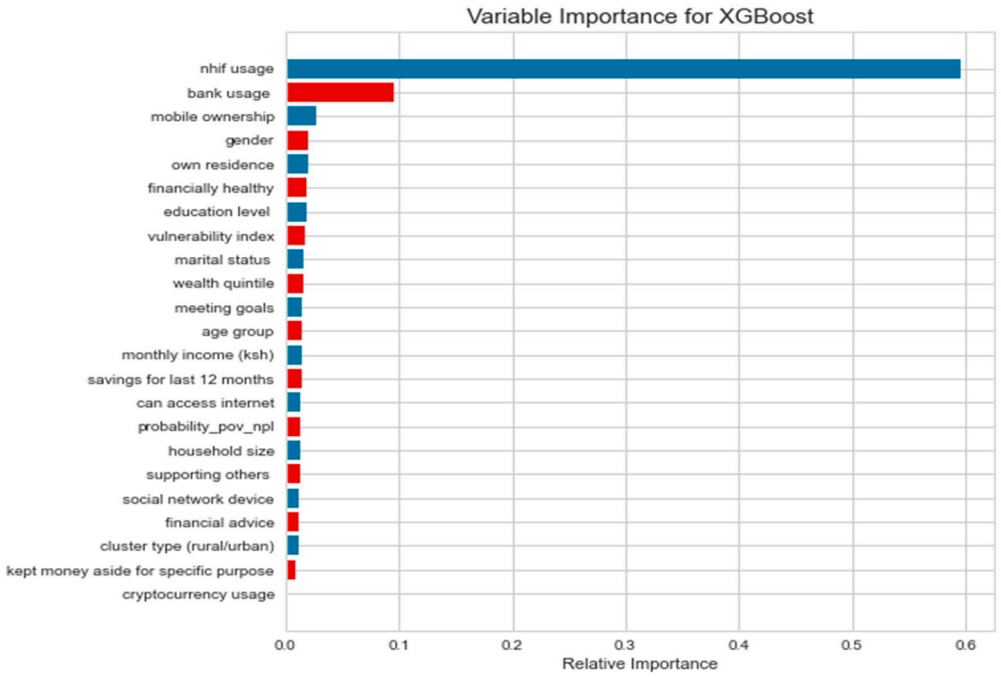

Figure 8. Feature importance from XGBoost.

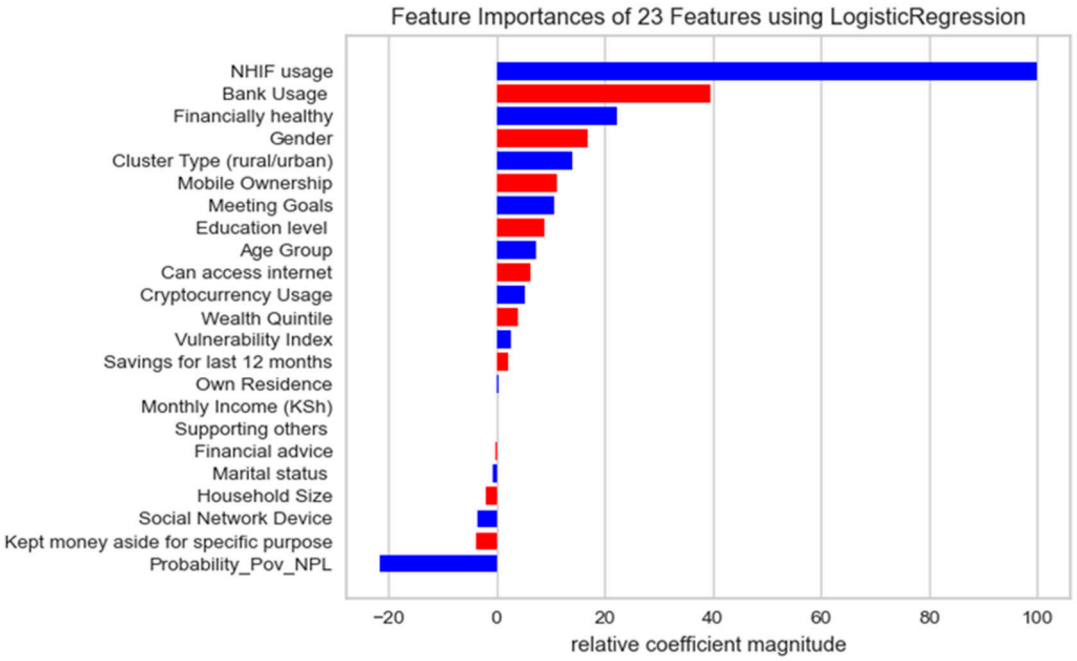

Figure 9. Feature importance from the logistic regression classifier.

Moreover, as Guven (2019) observed, the variable importance showed age as an important factor in influencing the uptake of pension. This concurs with Kitheka (2020), who postulated a positive association between the rise in age and uptake of pensions. This could imply that people become more conscious of the need to save for retirement as retirement approaches. Next, after age on importance was access to the Internet. This could imply that those who had internet access were more informed than those who did

not. As a result, they could make more informed choices in terms of saving and investing for retirement through membership in pension schemes. Moreover, cryptocurrency usage had the least importance. This implies that even usage of cryptocurrency had the least importance in determining whether one is enrolled in a pension scheme among the final features considered.

Figure 9 displays the variable importance in logistic regression. Unlike feature importance extracted from tree-based models, the variable importance from the logistic regression classifier shows both the size and the direction of the importance. Nevertheless, there is general congruence with regard to the size of the relative importance.

5. Conclusions

In conclusion, the results of this study suggest that ensemble tree-based models, specifically the random forest classifier, outperform both the decision tree classifier and logistic regression classifier in predicting pension uptake. The consistency of the results across unbalanced, up-sampled, and under-sampled data highlights the effectiveness of these models in this task. Furthermore, the superiority of the random forest classifier over XGBoost in precision, recall, F1 score, and accuracy, particularly for up-sampled data, indicates that this algorithm is the most robust model for pension uptake prediction in data of similar nature. These results suggest that policymakers and stakeholders in the pension sector should consider using the random forest classifier to optimize pension participation.

The study found that those who participated in the NHIF program were more likely to enroll in a pension plan. This highlights the significance of NHIF uptake in pension uptake. The study also supports the idea that promoting pension schemes could help in achieving universal health coverage and suggests that a combined approach of complementary NHIF policies and pension schemes may increase enrollment in both. The study found that monthly income, being banked, and support for others were the features that showed a positive relationship with pension uptake, suggesting that financial capacity is an important consideration for pension enrollment. The study also found that pension uptake increased with education level and age, implying that financial literacy and the realization of the need to save for retirement play a role in pension enrollment. Furthermore, the study found that access to the Internet was also a factor that influenced pension uptake, indicating that information plays a role in making informed choices about retirement savings. On the other hand, the study found that cryptocurrency usage had the least importance in determining pension enrollment among the factors considered.

Based on the findings of this study, several future directions could be considered to promote and optimize pension participation. Firstly, collaboration among various stakeholders, including regulators, pension providers, and related financial institutions, is needed to increase awareness and facilitate enrollment in pension schemes. The pension participation programs should aim to promote gender equality and empower both rural and urban dwellers. Additionally, financial education programs should be developed to enhance citizens' financial literacy and capacity, particularly for those with lower income and education levels. Furthermore, efforts should be made to improve access to information and technology, as the study found that internet access influenced pension uptake. Finally, future studies could explore the spatiotemporal aspects of pension uptake.

Author Contributions: N.K.Y., acquired the data, analyzed, and prepared the draft manuscript. J.K., as a principal advisor for N.K.Y., advised and supervised the data acquisition, analysis, and write up of the manuscript. J.N., advised and supervised the analysis and the write up of the manuscript. All authors have read and agreed to the published version of the manuscript.

Funding: The APC was funded by the African Center of Excellence in Data Science and the University of Rwanda.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data used in the study are available upon request from KNBS at https://www.knbs.or.ke/?wpdmpro=2019-finaccesshousehold-survey (accessed on 12 March 2021).

Acknowledgments: The authors thank FinAccess, FSD, and KNBS for publicly providing the data used in the study.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A

Appendix A.1. Relative Feature Importance

Table A1. Feature importance under the random forest classifier.

Covariate	Importance
NHIF usage	0.2583
Monthly income (KSh)	0.0965
Bank usage	0.0931
Probability_Pov_NPL (Poverty vulnerability)	0.0876
Supporting others	0.0510
Education level	0.0476
Household size	0.0439
Age group	0.0402
Can access internet	0.0351
Financial advice	0.0313
Financially healthy	0.0268
Wealth quintile	0.0245
Own residence	0.0235
Marital status	0.0201
Savings for last 12 months	0.0194
Gender	0.0189
Meeting goals	0.0157
Cluster type (rural/urban)	0.0150
Vulnerability index	0.0148
Kept money aside for specific purpose	0.0138
Social network device	0.0128
Mobile ownership	0.0095
Cryptocurrency usage	0.0003

Appendix A.2. Overview of the Dataset

Table A2. Dataset statistics.

Number of variables	30
Number of observations	8669
Missing cells	0
Missing cells (%)	0.0%
Duplicate rows	0
Duplicate rows (%)	0.0%
Total size in memory	2.0 MiB
Average record size in memory	240.0 B
Categorical	15
Boolean	11
Numerical	4
Categorical	15

Appendix A.3. Hyperparameters Tuned

Table A3. Hyperparameter tuned for the random forest classifier.

Parameter	Range	Optimal Value
n_estimators	[40 to 200, interval of 10]	110
max_features	[auto, sqrt, log2]	auto
min_samples_split	[2, 4, 6, 8]	2
Bootstrap	[true, false]	True

Table A4. Hyperparameter tuned for the XGBoost classifier.

Parameter	Range	Optimal Value
n_estimators	[500 to 1500]	1000
max_depth	[auto, sqrt, log2]	Auto
max_features	[0.2 to 1]	0.9
Gamma	[0.1 to 1]	0.1

Table A5. Hyperparameter tuned for the decision tree classifier.

Parameter	Range	Optimal Value
Splitter	[best, random]	Best
max_depth	[none, 1,2,3,4,5]	1
min_samples_split	[1,2,3, 5,6,7]	2
Criterion	[Gini, entropy]	Entropy

Appendix A.4. Description of Features

Table A6. Description of features.

Feature	Explanation	Range of Values
NHIF usage	NHIF usage	Yes, no
Monthly income (KSh)	Average monthly income in currency value (shillings)	In currency values
Bank usage	Having a formal bank account	Yes, no
Probability_Pov_NPL	Poverty vulnerability	0 to1 (continuous)
Supporting others	Amount of support to others	In currency values
Education level	Highest level of education level attained	None, primary, secondary, tertiary
Household size	Household size	Integer values (from 1 to 21)
Age group	Age group of the respondent	18–25 yrs, 26–35 yrs, 36–45 yrs, 45–55 yrs, >55 yrs
Can access internet	Can access internet	Yes, no, refused to answer
Financially healthy	Financially healthy	Yes, no
Wealth quintile	Wealth quintile level	Lowest, second lowest, middle, Second highest, highest
Own residence	Owning place of residence	Yes, no
Marital status	Marital status	Single/never married, married/living with partner, widowed
Savings for last 12 months	Average savings in currency for last 12 months	Yes, no
Gender	Gender of the respondent	Male, female
Meeting goals	Able to meet own financial goals	Yes, no
Cluster type (rural/urban)	Nature of residence (rural/urban)	Rural, urban
Vulnerability index	Vulnerability index	0 to1 (continuous)
Kept money aside for specific purpose	Kept money aside for specific purpose	Yes, no
Social network device	Having a social network device	Yes, no
Mobile ownership	Owning a mobile phone	Yes, no
Cryptocurrency usage	Using cryptocurrency	Yes, no

Appendix A.5. Figures on Trends in Pension Uptake

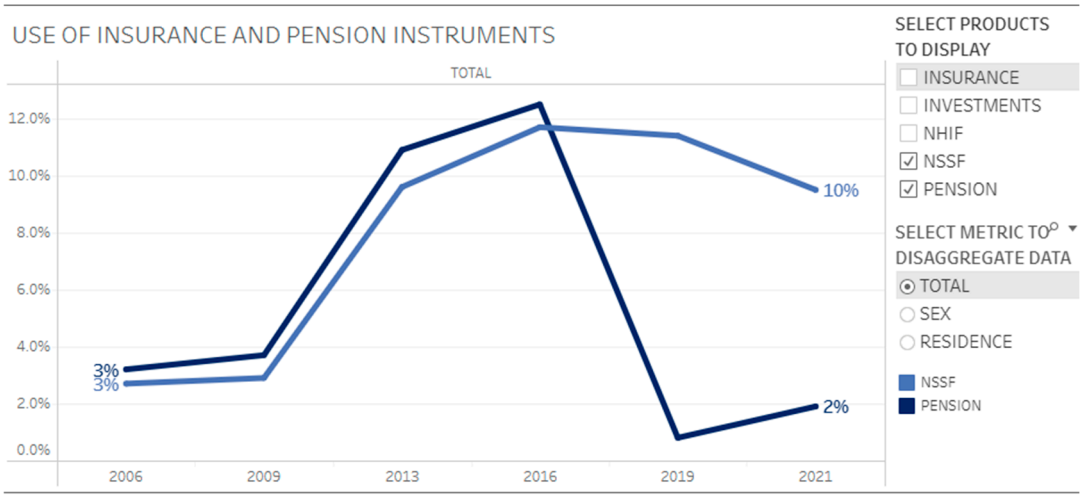

Figure A1. NSSF and other pension uptake over time. Source: FinAccess Kenya (2022).

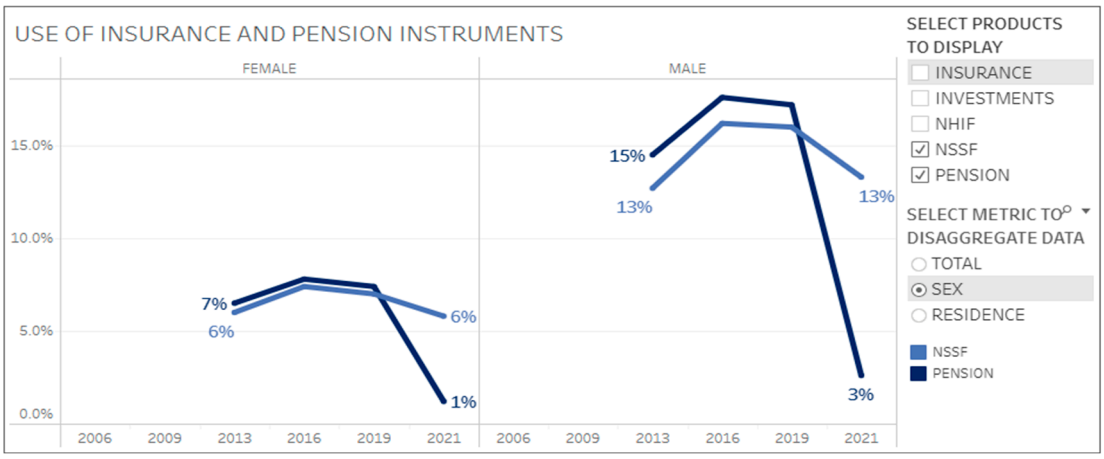

Figure A2. NSSF and other pension uptake by sex. Source: FinAccess Kenya (2022).

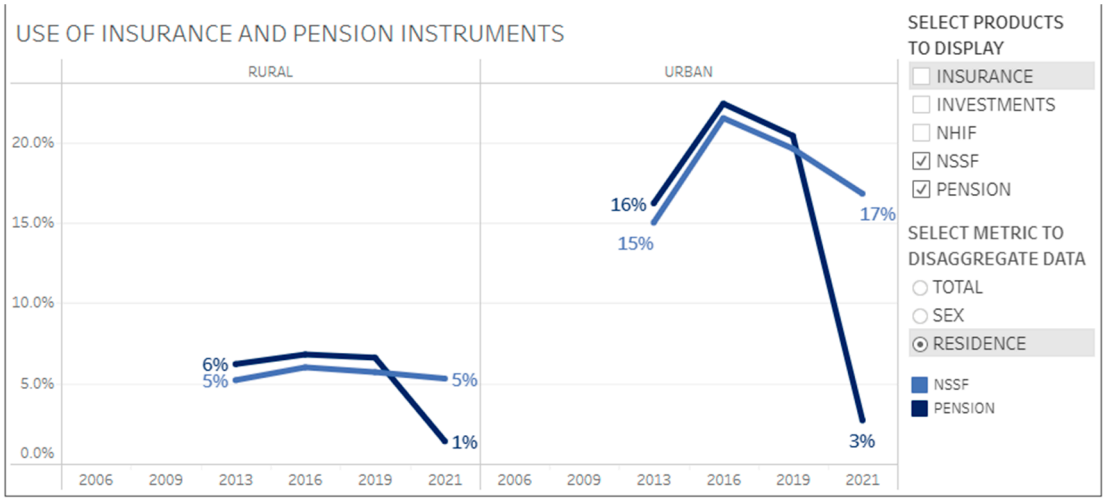

Figure A3. NSSF and other pension uptake by residence. Source: FinAccess Kenya (2022).

Appendix A.6. Correlogram and Correlation

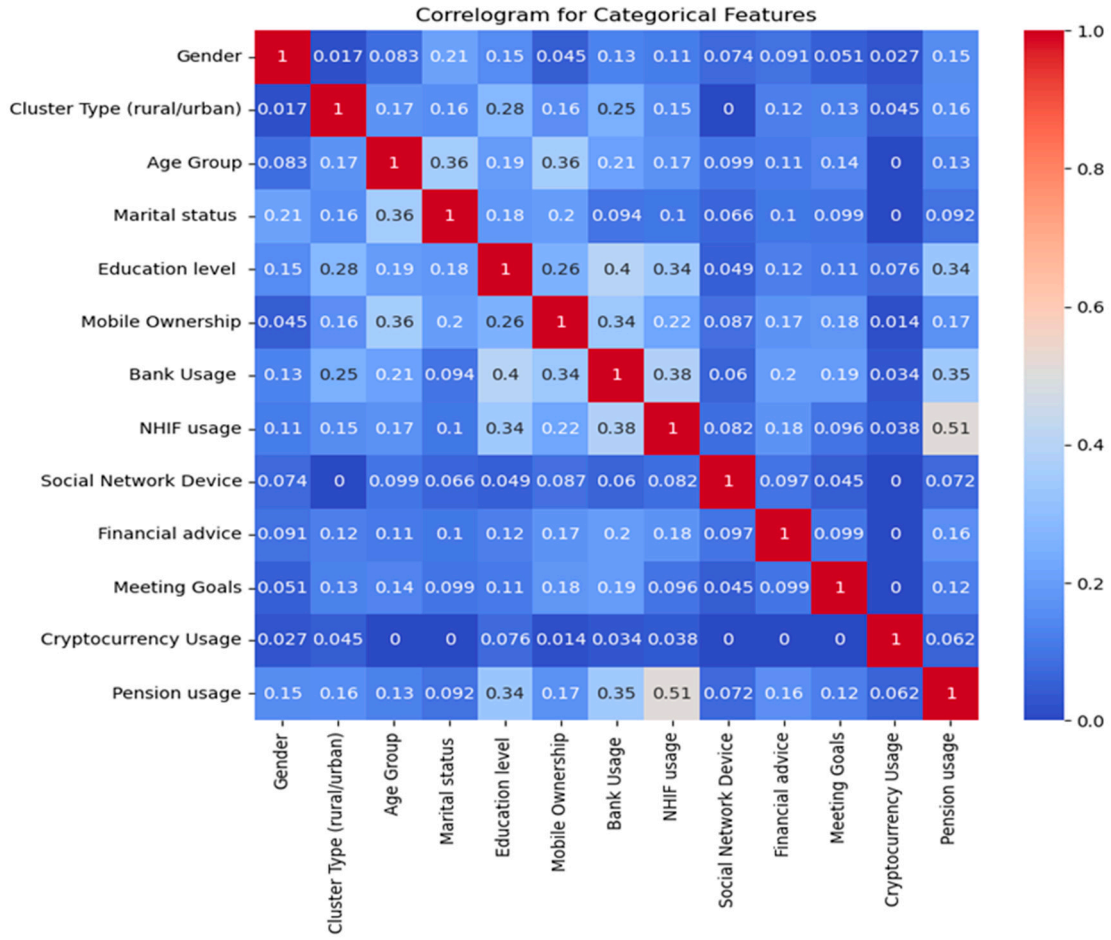

Figure A4. Correlogram for categorical features.

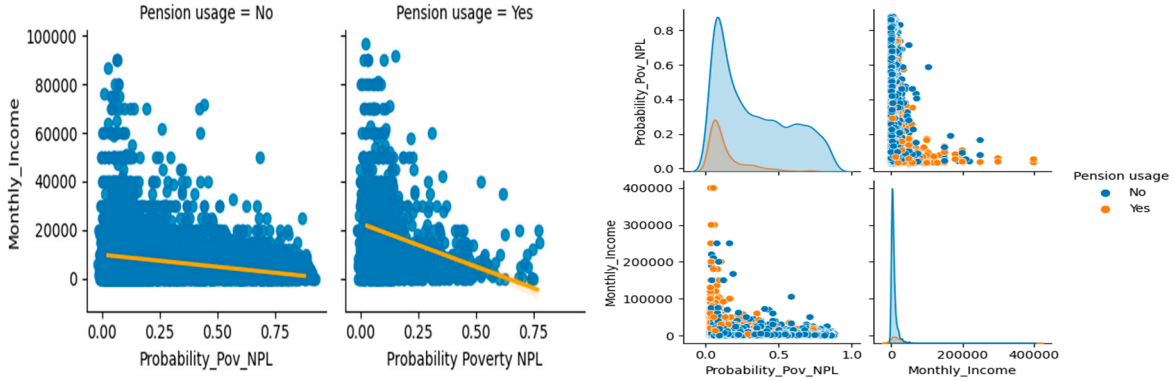

Figure A5. Correlation between poverty vulnerability, monthly income, and pension participation.

Note

[1] Tools and environment of the analysis: Python version 3.9.1 was the tool of choice used in the analysis, training of the models, and validation testing, as well as in performance metric comparison. Python was used because it was malleable to the analysis that was conducted. Moreover, the various libraries that Python has made the preprocessing, model training, and testing easier. Jupyter Notebook was the environment of choice for its simple interface.

References

Achoki, Tom, Molly K. Miller-Petrie, Scott D. Glenn, Nikhila Kalra, Abaleng Lesego, Gladwell K. Gathecha, and Uzma Alam. 2019. Health Disparities across the Counties of Kenya and Implications for Policy Makers, 1990–2016: A Systematic Analysis for the Global Burden of Disease Study 2016. *The Lancet Global Health* 7: e81–e95. [CrossRef] [PubMed]

Asuming, Patrick Opoku, Lotus Gyamfuah Osei-Agyei, and Jabir Ibrahim Mohammed. 2019. Financial Inclusion in Sub-Saharan Africa: Recent Trends and Determinants. *Journal of African Business* 20: 112–34. [CrossRef]

Balasuriya, Jiayi, and Yu Yang. 2019. The Role of Personality Traits in Pension Decisions: Findings and Policy Recommendations. *Applied Economics* 51: 2901–20. [CrossRef]

Barboza, Flavio, Herbert Kimura, and Edward Altman. 2017. Machine Learning Models and Bankruptcy Prediction. *Expert Systems with Applications* 83: 405–17. [CrossRef]

Barrera Ferro, David, Sally Brailsford, Cristián Bravo, and Honora Smith. 2020. Improving Healthcare Access Management by Predicting Patient No-Show Behaviour. *Decision Support Systems* 138: 113398. [CrossRef]

Belhadi, Amine, Sachin S. Kamble, Venkatesh Mani, Imane Benkhati, and Fatima Ezahra Touriki. 2021. An Ensemble Machine Learning Approach for Forecasting Credit Risk of Agricultural SMEs' Investments in Agriculture 4.0 through Supply Chain Finance. *Annals of Operations Research*. [CrossRef]

Bentéjac, Candice, Anna Csörgő, and Gonzalo Martínez-Muñoz. 2021. A Comparative Analysis of Gradient Boosting Algorithms. *Artificial Intelligence Review* 54: 1937–67. [CrossRef]

Bett, Nicholas, Juma Kasozi, and Daniel Ruturwa. 2022. Temporal Clustering of the Causes of Death for Mortality Modelling. *Risks* 10: 99. [CrossRef]

Boado Penas, Maria Del Carmen, Rocha Salazar, Jose de Jesús, Rocha Salazar, and Jose de Jesus. 2019. Scoring and Prediction of Early Retirement Using Machine Learning Techniques: Application to Private Pension Plans. *Anales Del Instituto de Actuarios Españoles* 25: 119–45.

Bouri, Elie, Konstantinos Gkillas, Rangan Gupta, and Christian Pierdzioch. 2021a. Forecasting Realized Volatility of Bitcoin: The Role of the Trade War. *Computational Economics* 57: 29–53. [CrossRef]

Breiman, Leo. 2001. Random Forests. *Machine Learning* 45: 5–32. [CrossRef]

Central Bank of Kenya, FSD Kenya, and Kenya National Bureau of Statistics. 2019a. *FinAccess Household Survey 2019*. Nairobi: FinAccess.

Central Bank of Kenya, Kenya National Bureau of Statistics, and FSD Kenya. 2019b. *The 2019 FinAccess Household Survey*. Data. Nairobi: FinAccess. Available online: https://fsdkenya.org/publication/finaccess2019/ (accessed on 17 November 2021).

Chandrashekar, Girish, and Ferat Sahin. 2014. A Survey on Feature Selection Methods. *Computers & Electrical Engineering* 40: 16–28. [CrossRef]

Cobb-Clark, Deborah A., Sarah C. Dahmann, Daniel A. Kamhöfer, and Hannah Schildberg-Hörisch. 2022. The Predictive Power of Self-Control for Life Outcomes. *Journal of Economic Behavior & Organization* 197: 725–44. [CrossRef]

Consumer Options Ltd. 2019. Private Sector Analysis Survey on Saving for Retirement. NAIROBI. Available online: https://www.rba.go.ke/download/private-sector-analysis-survey-on-saving-for-retirement-co/# (accessed on 3 November 2020).

Diana, Alex, Jim E. Griffin, Jaideep S. Oberoi, and Ji Yao. 2019. *Machine-Learning Methods for Insurance Applications—A Survey*. Schaumburg: Society of Actuaries.

Dixon, Matthew F., Igor Halperin, and Paul Bilokon. 2020. *Machine Learning in Finance*. Cham: Springer Nature, vol. 1170. [CrossRef]

FinAccess Kenya. 2022. Use of Insurance and Pensions. Dashboard. Available online: https://finaccess.knbs.or.ke/usage/use-of-insurance-and-pensions (accessed on 20 March 2023).

Guerrero, Maribel, Francisco Liñán, and F Rafael Cáceres-Carrasco. 2021. The Influence of Ecosystems on the Entrepreneurship Process: A Comparison across Developed and Developing Economies. *Small Business Economics* 57: 1733–59. [CrossRef]

Guven, Melis. 2019. *Extending Pension Coverage to the Informal Sector in Africa*. Social Protection and Jobs Discussion Paper, No. 1933. Washington, DC: World Bank Group. [CrossRef]

Hadad, Elroi, Stanislav Dimitrov, and Jivka Stoilova-Nikolova. 2022. Development of Capital Pension Funds in the Czech Republic and Bulgaria and Readiness to Implement PEPP. *European Journal of Social Security* 24: 342–60. [CrossRef]

Kibona, Shadrack Elia. 2020. Determinants of Pension Uptake in the Informal Sector of Tanzania. *Huria Journal of the Open University of Tanzania* 27: 17–28.

Kipkogei, Francis, Ignace H. Kabano, Belle Fille Murorunkwere, and Nzabanita Joseph. 2021. Business Success Prediction in Rwanda: A Comparison of Tree-Based Models and Logistic Regression Classifiers. *SN Business & Economics* 1: 101. [CrossRef]

Kitheka, Paul. 2020. *Factors Influencing Intent of Uptake of Retirement Pension and Provident Scheme Plans in the Informal Sector in Nairobi County*. Nairobi: Strathmore University. Available online: http://hdl.handle.net/11071/12070 (accessed on 17 November 2021).

Künzler, Daniel. 2016. Social Security Reforms in Kenya: Towards a Workerist or a Citizenship-Based System? *International Social Security Review* 69: 67–86. [CrossRef]

Kutrani, Huda, Saria Eltalhi, and Naeima Ashleik. 2021. Predicting Factors Influencing Survival of Breast Cancer Patients Using Logistic Regression of Machine Learning. Paper presented at the 7th International Conference on Engineering & MIS 2021, Almaty, Kazakhstan, October 11–13, pp. 1–6.

Kwena, Rose Musonye, and John A. Turner. 2013. Extending Pension and Savings Scheme Coverage to the Informal Sector: Kenya's Mbao Pension Plan. *International Social Security Review* 66: 79–99. [CrossRef]

Lades, Leonhard K., Mark Egan, Liam Delaney, and Michael Daly. 2017. Childhood Self-Control and Adult Pension Participation. *Economics Letters* 161: 102–4. [CrossRef]
Levantesi, Susanna, and Giulia Zacchia. 2021. Machine Learning and Financial Literacy: An Exploration of Factors Influencing Financial Knowledge in Italy. *Journal of Risk and Financial Management* 14: 120. [CrossRef]
Li, Yaqi, Chun Yan, Wei Liu, and Maozhen Li. 2016. Research and Application of Random Forest Model in Mining Automobile Insurance Fraud. Paper presented at the 2016 12th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery, ICNC-FSKD 2016, Zhangjiajie, China, August 13–15, pp. 1756–61. [CrossRef]
Lin, Weiwei, Ziming Wu, Longxin Lin, Angzhan Wen, and Jin Li. 2017. An Ensemble Random Forest Algorithm for Insurance Big Data Analysis. *IEEE Access* 5: 16568–75. [CrossRef]
Mare, Codruţa, Daniela Manaţe, Gabriela-Mihaela Mureşan, Simona L. Dragoş, Cristian M. Dragoş, and Alexandra-Anca Purcel. 2022. Machine Learning Models for Predicting Romanian Farmers' Purchase of Crop Insurance. *Mathematics* 10: 3625. [CrossRef]
Mutai, C. K., P. E. McSharry, I. Ngaruye, and E. Musabanganji. 2021. Use of machine learning techniques to identify HIV predictors for screening in sub-Saharan Africa. *BMC Medical Research Methodology* 21: 159. [CrossRef]
Pörtner, Hans-O., Debra C. Roberts, Helen Adams, Carolina Adler, Paulina Aldunce, Elham Ali, and Rawshan Ara Begum. 2022. *Climate Change 2022: Impacts, Adaptation and Vulnerability*. Geneva: IPCC.
Rajan, Irudaya S., Pooja Batra, Reddy Sai Shiva Jayanth, and Tharatha Moolayil Sivadasan. 2023. Understanding the Multifaceted Impact of COVID-19 on Migrants in Kerala, India. *Development Policy Review* 41: e12636. [CrossRef] [PubMed]
Renault, Thomas. 2020. Sentiment Analysis and Machine Learning in Finance: A Comparison of Methods and Models on One Million Messages. *Digital Finance* 2: 1–13. [CrossRef]
Riumallo-Herl, Carlos, and Emma Aguila. 2019. The Effect of Old-Age Pensions on Health Care Utilization Patterns and Insurance Uptake in Mexico. *BMJ Global Health* 4: 1–10. [CrossRef]
Saroj, Rakesh Kumar, and Madhu Anand. 2021. Environmental Factors Prediction in Preterm Birth Using Comparison between Logistic Regression and Decision Tree Methods: An Exploratory Analysis. *Social Sciences & Humanities Open* 4: 100216. [CrossRef]
Serrano, Antonio Sánchez, and Tuomas Peltonen. 2020. *Pension Schemes in the European Union: Challenges and Implications From Macroeconomic and Financial Stability Perspectives*. No. 17. ESRB Occasional Paper Series. Frankfurt am Main: ESRB. [CrossRef]
Shalev-Shwartz, Shai, and Shai Ben-David. 2014. *Understanding Machine Learning: From Theory to Algorithms*. New York: Cambridge University Press.
Unnikrishnan, Vidhya, and Katsushi S. Imai. 2020. Does the Old-Age Pension Scheme Improve Household Welfare? Evidence from India. *World Development* 134: 105017. [CrossRef]
Yego, Nelson Kemboi, Juma Kasozi, and Joseph Nkurunziza. 2021. A Comparative Analysis of Machine Learning Models for the Prediction of Insurance Uptake in Kenya. *Data* 6: 116. [CrossRef]