Digital Financial Advisor Using Random Forest Regressor

Bidisha Amatya[1,*], Prasanna Shakya[2], Prinska Maharjan[3], Sajal Maharjan[4], Ashok GM[5]
[1,2,3,4,5]Himalaya College of Engineering, Tribhuvan University (TU), Lalitpur, Nepal
*Corresponding author: bidisha.amatya@gmail.com

Abstract

This study presents the Digital Financial Advisor, a webbased application designed to provide personalized invest recommendations by analyzing user specific financial factors such as age, financial knowledge, risk tolerance, investment time horizon, and financial goals. The system uses the Random Forest Regressor algorithm as the core methodology to generate optimal portfolio allocations across diverse asset classes like stocks, fixed deposits, SIPs, bonds, and commodities. The platform features interactive pie chart that makes the investment strategies easier to understand. Initial testing demonstrated system responsiveness and effective port- folio distribution. In conclusion, the system bridges the gap between expert financial advice and everyday users. Future enhancements include integrating real-time financial data for dynamic recommendations and expanding feature sets to incorporate additional financial factors, ensuring more comprehensive and accurate investment guidance.

Keywords: Digital Financial Advisor, Investment Recommendations, Financial Factors, Random Forest Regressor, Portfolio Allocation, Asset Classes

1. Introduction

In today's dynamic financial landscape, making informed investment decisions is a challenge for many individuals, particularly those without access to expert guidance. Many rely on limited knowledge or generic online resources, which often fail to consider their specific financial circumstances. Existing financial advisory applications, such as Robinhood, Acorns, and Betterment, provide valuable tools for portfolio management and in- vestment tracking, but they rarely deliver personalized recommendations tailored to an individual's financial literacy, risk tolerance, or long-term goals. In Nepal, where financial literacy remains comparatively low and most available apps focus on stock or IPO management, individuals are frequently left to make investment choices without clear, data-driven guidance.

Traditional investment advisory models typically rely on rule-based optimization or static allocation strategies. While these provide simple heuristics, they lack adaptability to evolving financial behaviors and diverse asset classes. In contrast, machine learning methods such as ensemble models (e.g., Random Forest) have shown superior performance over traditional econometric models in handling nonlinear financial data [3]. Compared to deep learning approaches, Random Forest requires fewer data samples and offers greater interpretability, making it well- suited for personalized financial advisory applications.Among these methods, Random Forests have also been widely used to rank feature importance in financial models, demonstrating robustness against overfitting and facilitating interpretability [4].

Another key challenge is the scarcity of high- quality, labeled financial datasets in local contexts like Nepal.To address data scarcity and privacy concerns, we opted for synthetic data generation, a method increasingly adopted in financial modeling contexts [6]. By generating user profiles that reflect key financial factors such as age, financial knowledge, risk tolerance, time horizon, and investment goals, our system ensures realistic input data while maintaining user privacy.

The objective of this project is therefore to develop a web-based Digital Financial Advisor that computes portfolio allocation ratios across diverse asset classes using a Random Forest Regressor. By combining robust machine learning methods, synthetic dataset generation, and personalized recommendation logic, the system aims to bridge the gap between expert financial advice and the needs of everyday users in Nepal.

2. Literature Review

In recent years, technological advancements have significantly transformed the financial services industry, leading to the development of automated and intelligent portfolio management tools. Investment portfolio management applications aim to help individuals make informed decisions about asset allocation, balancing risk and return.

In developing the portfolio for this project, we synthesized data based on a commonly cited rule of thumb which has helped simplify asset allocation. According to this principle, individuals should hold a percentage of stocks equal to 100 minus their age [5]. In addition to this rule, we used a set of guidelines derived from Vanguard's investment principles. The process begins with identifying financial goals to establish a clear purpose for investment, such as retirement, education, or asset growth. Next, it evaluates risk tolerance, tailoring asset allocation to align with the investor's comfort level regarding market fluctuations. The time horizon for investments is then assessed to adjust the portfolio's aggressiveness or conservativeness appropriately [8].

The fast expansion of automated financial advisory systems, also referred to as "robo-advisors," has extended the evolution of the investment world. These systems make algorithmic financial planning accessible even without human interven-

Article History:
Submission: 2025-05-30, Revision: 2025-09-10, Acceptance: 2025-11-11

tion. Utilizing Modern Portfolio Theory (MPT) and algorithms such as mean-variance optimization, robo-advisors in- vest, manage, and optimize portfolios based on user requirements including risk tolerance and investment objectives [2].

Wealthfront is a robo-investment service that uses Mean-Variance Optimization , the building block of Modern Portfolio Theory, to determine the optimal asset class portfolio. It produces efficient portfolios with maximum return for given amounts of risk or minimum risk for given expected returns, generating the efficient frontier. Wealthfront is most appropriate for customers who are willing to have hands-off, continuous management [9].

Betterment is another leading robo-advisory platform that builds diversified portfolios of low-expense index funds aligned with the investor's goals, risk tolerance, and investment time frame. After users input details about their aims and risk tolerance, Betterment creates a tailor-made asset allocation strategy based on MPT principles. Its automation advantages include tax-loss harvesting and rebalancing, making it a favorite for easy, low- expense investment management [1].

In Nepal, Money Mitra is an emerging fintech application that provides basic financial tracking, mutual fund investment, and stock market data access. However, it focuses mainly on execution rather than advisory, lacking personalized investment guidance or portfolio optimization. This highlights the need for applications that not only facilitate transactions but also educate users and provide tailored advice, especially for those with limited financial literacy [7].

To enhance portfolio management, algorithms like Random Forest (RF) can complement the strategies used by Wealthfront and Betterment. RF leverages historical data to refine recommendations, improving the accuracy and stability of investment decisions. The RF model uses bootstrap sampling to generate training subsets, builds multiple decision trees, and aggregates predictions through majority voting. This approach prioritizes stocks with higher stability and lower prediction risk, ensuring robust recommendations [3].

Studies emphasize the necessity of tools that simplify investment processes for users with limited financial literacy. Applications that provide predefined investment portfolio options, complemented by detailed insights into their risk-return trade-offs, can empower users to make informed decisions. By integrating local financial data, such tools not only enhance user accessibility but also contribute to the broader financial inclusion agenda by encouraging investment participation [7].

3. Methodology

The methodology follows an approach that includes four main iterations: data preprocessing, model training, model tuning and web app development.

3.1 Dataset Generation

A key challenge faced during this project was the lack of publicly available, labeled financial advisory datasets in the

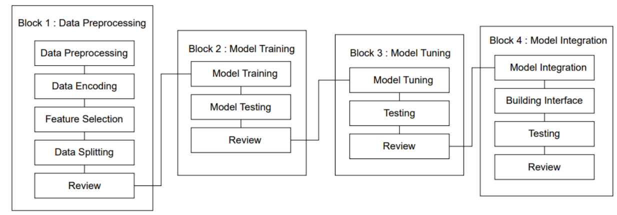

Figure 1. Development Methedology

Nepalese context. Most existing datasets are either proprietary or focused on specific markets such as the U.S. or Europe. Additionally, concerns regarding privacy and user sensitivity in financial data limit the feasibility of directly collecting real user data. To address these issues, we generated a synthetic dataset, a method increasingly adopted in financial modeling research. ([6] This ensured a sufficient volume of training data while maintaining data privacy and relevance to local financial conditions. The synthetic dataset was created using proportions derived from established financial heuristics and portfolio management guide- lines. For instance, the widely cited "100 mi- nus age" rule suggests that individuals should hold a percentage of stocks equal to 100 Age [5]. Additional portfolio allocation principles were adapted from Vanguard's investment guidelines, which emphasize balancing risk tolerance, time horizon, and financial goals [8]. By combining these heuristics with randomized sampling, we generated realistic yet diverse user profiles.

1) Dataset Characteristics

The final data set consisted of approximately 20,000 records. Each record represents a unique investor profile with six input features and six output targets:

- Input Features
  - Age: Integer values ranging from 18 to 70 years.
  - Financial Knowledge: Encoded as Low (0), Medium (1), High (2).
  - Risk Appetite: Encoded as Low (0), Medium (1), High (2).
  - Comfort with Fluctuations: Encoded as Low (0), Medium (1),High (2).
  - Investment Horizon: Categorized as Short (1–3 years), Medium (3–10 years), Long (10+ years).
  - Investment Goal: One-hot encoded as Wealth Growth, Stability, or Retirement.
- Target Variables (Portfolio Allocation Ratios) Percentage allocations across six asset classes:
  - Stocks
  - Fixed Deposits (FD)
  - Systematic Investment Plans (SIPs)
  - Bonds/Debentures
  - Gold/Silver
  - Cash

The allocation ratios were generated such that they always sum to 100%, with adjustments made according to the user profile. For example, younger users with high risk tolerance received larger allocations in stocks and SIPs, while older users with conservative preferences were assigned higher proportions in fixed deposits and bonds. This rule-driven synthetic approach ensured realistic portfolio patterns that align with widely accepted financial advisory practices.

3.2 Data Preprocessing

Data Preprocessing plays a crucial role in preparing high-quality input data for the Random Forest model. To effectively train the system, synthetic data was generated based on the Standard Principles of Financial and Portfolio Management, ensuring alignment with user profiles and financial goals. Investment portfolios were categorized using parameters such as financial knowledge, age, investment time horizon, risk appetite, comfort with fluctuations, and investment objectives. Since these inputs are often qualitative, Data Encoding was applied to convert categorical values into a structured numeric format suitable for machine learning. Specifically, Ordinal/Label Encoding was used to map qualitative inputs to numeric values, such as Risk Appetite (”Low”$\rightarrow$ 0, ”Medium” $\rightarrow$ 1, ”High” $\rightarrow$ 2) and one-hot encoding was used to map qualitative input i.e. Investment Goals as (”Wealth Growth”$\rightarrow$ [1, 0, 0], “Stability” $\rightarrow$ [0, 1, 0], ”Retirement”$\rightarrow$ [0, 0, 1]). To enhance model performance, the preprocessing stage also involved Data Cleaning and Feature Engineering. The next step was Feature Selection, where the most relevant input features were identified to optimize model accuracy and computational efficiency.

3.2.1 Model Training

The Digital Financial Advisor Model was trained and tested using a Random Forest Regression algorithm. This machine learning approach was chosen for its robustness in handling non-linear relationships and its ability to capture complex interactions between input parameters.

Training the Model

The Random Forest Regression model was trained using multiple decision trees, where each tree learned from different subsets of the dataset. The model aggregated predictions from all trees, reducing overfitting and improving generalization. The implementation of random forest regressor is explained below:

1) **Bootstrap Sampling (Selecting Data):** The dataset is randomly sampled with replacement to create multiple training sub- sets. Each tree is trained on a slightly different dataset.
2) **Feature Selection:** In Random Forest, feature selection occured automatically during model training. At each decision split, the algorithm considered only a random subset of features instead of all available ones, selecting the feature that provided the best split. This randomness reduced the dominance of strong predictors, promote diversity among trees, and prevented overfitting. Additionally, Ran- dom Forest provided feature importance scores, based on either the reduction in impurity or the impact on prediction accuracy, which helped identify the most influential variables for the model.This dual approach ensured that the model achieved both predictive accuracy and practical applicability in financial advisory contexts. .
3) **Splitting Criterion:** Since this is a regression task, the Mean Squared Error (MSE) is used to determine the best split: $\text{MSE}(N) = \frac{1}{|N|}\sum_{i\in N}(y_i - \bar{y}_N)^2$
The feature that results in the largest reduction in MSE is selected for the split.
4) **Combining Multiple Trees:** Once multiple decision trees are trained, their outputs are averaged to produce a final portfolio allocation recommendation.The detailed explanation of how the model is trained is further explained below:
   - The dataset was loaded, with input features (X) including user attributes such as Age, Financial Knowledge, Risk Appetite, Comfort with Fluctuation, Time Horizon, and Investment Goal.
   - The target variables (y) represented the percentage allocations across six asset classes: Stocks, Fixed Deposits (FD), Systematic Investment Plans (SIPs), Bonds/Debentures, Gold/Silver, and Cash.
   - The data was split into training (80%) and testing (20%) subsets to evaluate the model’s predictive performance.
   - A Random Forest Regressor was configured using 100 Decision Trees to ensure a balance between accuracy and computational efficiency and Mean Squared Error (MSE) as the loss function to minimize prediction errors.

Model Evaluation

A Random Forest Regression model was employed for its robustness in handling complex financial data. The dataset included user attributes Age, Financial Knowledge, Risk Appetite, Comfort with Fluctuation, Time Horizon, and Investment Goal as input features, with target variables representing percentage allocations across six asset classes: Stocks, Fixed Deposits, SIPs, Bonds/Debentures, Gold/Silver, and Cash. Data was split 80:20 for training and testing, using 100 decision trees and Mean Squared Error (MSE) as the loss function.

Performance metrics

- **MSE:**
  Measures average squared prediction errors; lower values indicate higher accuracy
- R^2 **Score:**
  Indicates proportion of variance explained by the model; values closer to 1 denote better fit.
- **Feature Importance:**
  Highlights key predictors such as Age, Risk Appetite, Comfort with Fluctuation, and Financial Knowledge, enhancing model interpretability and transparency in investment recommendations.

3.2.2 Model Tuning

Model Tuning involved optimizing the Random Forest Regression model to enhance its predictive performance and generalizability. This was achieved through hyperparameter tuning, feature engineering, outlier removal, and model evaluation techniques.

Data Preprocessing & Feature Engineering

To ensure robust training, several preprocessing steps were implemented:

- Outlier Removal: The Interquartile Range (IQR) method was applied to remove extreme values from the dataset, preventing skewed model behavior.
- Feature Scaling: The MinMaxScaler was used to normalize feature values between 0 and 1, ensuring the model treated all variables uniformly.

Hyperparameter Optimization A comprehensive hyperparameter tuning process was conducted using Randomized Search Cross-Validation (RandomizedSearchCV),which systematically explored various combinations of hyperparameters to enhance model performance. The optimized parameters included number of trees (n estimators) which was increased up to 1000 to improve robust- ness. Tree depth (max depth) was tuned to control model complexity. Minimum samples for split (min samples split) was adjusted to prevent overfitting on small samples. Mini- mum samples per leaf (min samples leaf) was set to optimize decision boundaries. Feature selection (max features) where different feature subsets (auto, sqrt, log2) were tested. Bootstrap sampling (bootstrap) was enabled and disabled to compare performance.

The best hyperparameters were identified through 10-fold cross-validation, ensuring model stability across different data subsets.

3.2.3 Model Tuning

Trained and tuned model was integrated in the digital financial advisor system, where web app was built using the Django framework for the backend, React for the frontend and SQLite for the database. It includes essen- tial modules such as user login, signup, and form-based queries for investment classifica- tion. Additionally, admin functionalities en- able application updates and secure access.

3.3 System Flow Diagram

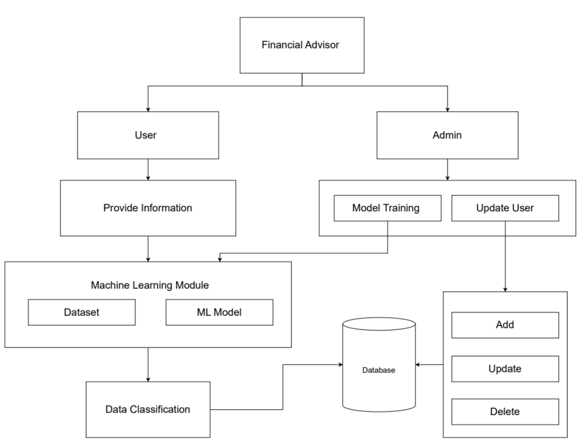

Figure 2. System Flow Diagram

This diagram illustrates the workflow of a financial advisory system, which comprises three primary components: the User, the Admin, and a Machine Learning Module. Users input their financial information into the system, while Admins manage tasks such as training the model and updating user data. The Ma- chine Learning Module processes this information using a combination of a dataset and a machine learning model to classify and analyze the data. The results help in making personalized financial recommendations. Additionally, a Database component stores and manages the user data, enabling the addition, updating, or deletion of information. This integrated process ensures the system can offer tailored financial advice based on users' specific financial details and needs.

4. Results and Analysis

4.1 Model Performance Evaluation

The Table 1 presents the model performance evaluation of random forest regressor model for Digital Financial Advisor system.

Table 1. Model Evaluation Metrics

Metric	Value
Mean Squared Error (MSE)	5.04
R-squared Score (R^2)	0.81
Variance	17.07

Mean Squared Error(MSE) measured the average squared difference between actual and predicted values. A lower MSE indicated high prediction accuracy and model reliability. The model achieved an MSE of 5.04, confirming minimal deviation from actual portfolio allocations and high prediction accuracy. The R^2score represented the proportion of variance in the target variable explained by the model, ranging from 0 to 1. Values closer to 1 indicate better model fit. The model obtained an R^2 score of 0.81, demonstrating its capability to capture complex relationships between user attributes and investment decisions. The model's variance was 17.07, which indicates a notable difference from the MSE of 5.04. This further confirms that the model was functioning effectively and not overfitting.

Scatter Plots

These scatter plots visualize the Actual vs. Predicted values

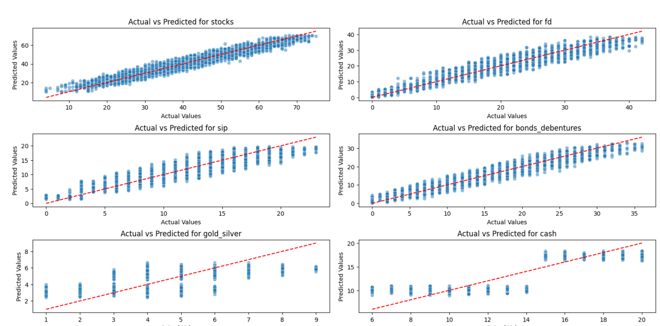

Figure 3. Scatter Plots

for portfolio allocations across six asset classes: Stocks, Fixed Deposits (FD), SIP, Bonds/Debentures, Gold/Silver, and Cash. In each plot, the X- axis represents the actual values from the test set. The Y-axis represents the values predicted by the Random Forest Regression model and red dashed line is the ideal line where predicted values exactly match actual values. The data points closely aligned with the red line showing a consistent pattern along the diagonal indicates high prediction accuracy. Model shows good prediction accuracy for all asset classes except cash which may possibly be due to data imbalance.

Feature Importance Analysis

To better understand the factors influencing the model's predictions, a feature importance analysis was conducted. This analysis highlighted which input features had the greatest impact on the decision-making process of the Random Forest Regression model for portfolio allocation. The bar

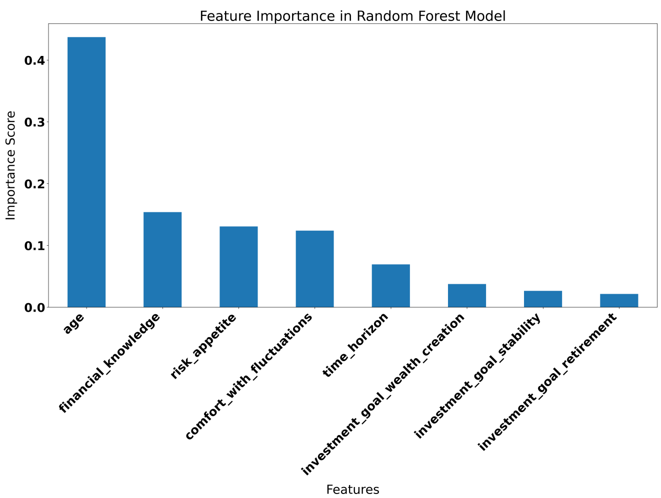

Figure 4. Feature Importance Bar Graph

graph illustrates the relative importance of input features used by the Random Forest regression model in the Digital Financial Advisor system. Each importance score reflects the extent to which a particular feature contributes to the model's portfolio allocation predictions. Among all input features, age emerges as the most influential, having the highest importance score. This indicates that age significantly impacts the model's investment recommendations. Features such as financial knowledge, risk appetite, and comfort with fluctuations show moderate importance, suggesting that they meaningfully affect how the model distributes investments across as- set classes. In contrast, investment goal and time horizon exhibit relatively lower importance scores, implying that their influence on the model's predictions is comparatively limited.

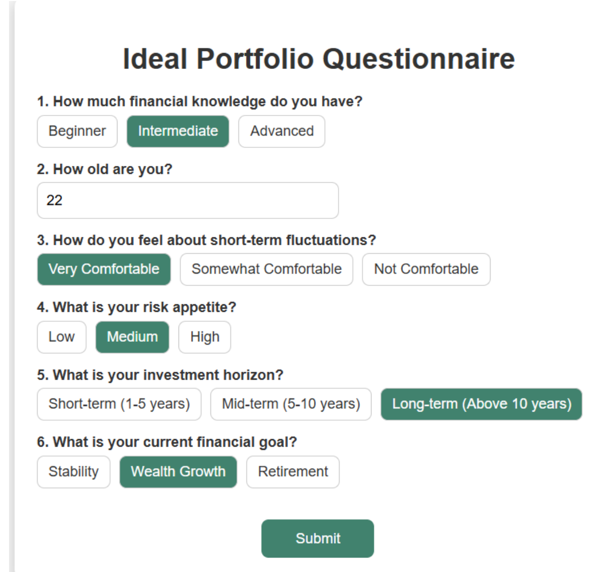

Figure 5. Ideal Portfolio Questionaire

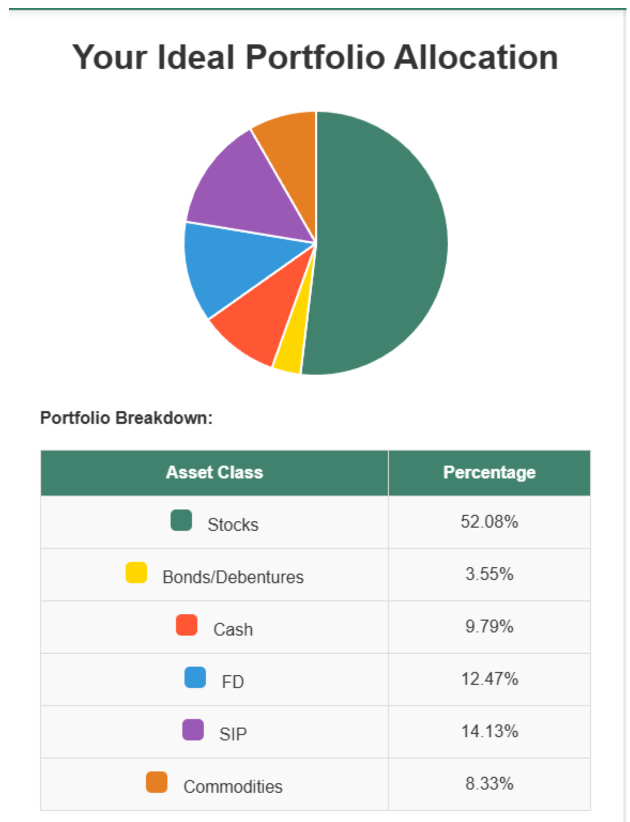

Figure 6. Recommended Portfolio

5. Discussion

The Digital Financial Advisor demonstrates a practical, user-friendly solution for personalized investment planning, especially valuable in Nepal's under-served financial landscape. The system uses the Random Forest Regressor to analyze user-specific factors such as age, financial knowledge, risk tolerance, time horizon, and goals and then delivers tailored portfolio recommendations across diverse asset classes. The reliability of the model is confirmed by performance metrics: even with limited local data, an MSE of 5.04 and an R^2 score of 0.81 show high accuracy and strong explanatory power (Section 4). Users of all literacy levels can understand complex financial strategies with the help of interactive pie charts.

Reliance on static user data and absence of real-time market integration are current drawbacks. For accuracy and adaptability, future research should focus on dynamic data input and improved user features. The project's main achievement is combining machine learning with accessible visualization for resource-limited environments, offering a scalable, affordable alternative to traditional or imported financial advising tools. The system empowers users to make informed investment decisions and provides valuable educational resources for financial literacy.

6. Conclusion

This research developed The Digital Financial Advisor system, achieving its goal of providing accurate, personalized investment recommendations using Random Forest Regression. Comprehensive evaluation confirmed high prediction accuracy and model interpretability. Efficient system integration ensured seamless component interaction, while a responsive, intuitive user interface delivered a smooth experience across devices. Certain challenges like data Imbalance, complexity in feature selecion and integration hurdles were encountered. This project laid a strong foundation for future enhancements to increase system efficiency and adaptability. One key improvement is incorporating real-time financial data for dynamic recommendations. Expanding the feature set to include additional financial factors is another enhancement. Finally, advanced machine learning models can be implemented to improve accuracy. Overall, the project demonstrated the potential of data-driven financial advisory systems, contributing to personalized investment planning and financial decision making.

Acknowledgment

We express our gratitude to Himalaya College of Engineering for providing the environment and resources necessary to complete this research. We are also deeply grateful to Er. Ashok GM, Head of Department, for his out- standing leadership and invaluable guidance, and to Er. Ramesh Tamang, our Project Co-ordinator, for his knowledgeable counsel and hardworking oversight. Their encouragement and support have been crucial to our work's successful conclusion.

References

[1] Betterment, "Betterment: Smarter investing for the future," 2024. [Online]. Available: https://www.betterment.com. [Accessed: Dec. 16, 2024].

[2] J. Bogle, "The future of investment management," *Journal of Portfolio Management*, vol. 40, no. 4, pp. 23–35, 2014.

[3] M. Chen, Y. Zhang, and X. Luo, "A quantitative investment model based on random forest and sentiment analysis," *Journal of Physics: Conference Series*, vol. 1575, no. 1, 012083, 2020.

[4] H. Gao, G. Kou, H. Liang, Y. Wang, and Y. Peng, "Machine learning in business and finance: A literature review and research opportunities," *Financial Innovation*, vol. 10, no. 1, p. 86, 2024, doi: 10.1186/s40854-024-00629-z.

[5] Investopedia, "Stock allocation rules," 2024. [Online]. Available: https://www.investopedia.com/articles/investing/062714/100-minus-your-age-outdated.asp. [Accessed: Dec. 30, 2024].

[6] JPMorgan Chase, "Synthetic data in financial modeling: Opportunities and challenges," Tech. Rep., 2023.

[7] K. Shrestha, "The state of fintech in Nepal: Challenges and opportunities for financial inclusion," *Nepal Journal of Economic Studies*, vol. 4, no. 2, pp. 45–58, 2023.

[8] Vanguard, "Investor education: Balancing risk and reward," 2024. [Online]. Available: https://investor.vanguard.com/. [Accessed: Dec. 16, 2024].

[9] Wealthfront, "Investment methodology," 2024. [Online]. Available: https://research.wealthfront.com/whitepapers/investment-methodology. [Accessed: Dec. 16, 2024].