Journal of Applied Informatics and Computing (JAIC)
Vol.9, No.6, December 2025, pp. 3130~3137
e-ISSN: 2548-6861

Binary Classification for Predicting the Investment Trends of The Younger Generation Based on Machine Learning

Weka Brilliant Jaya Oktana[1]*, Ucta Pradema Sanjaya [2]*
* Teknik Informatika, Universitas Ngudi Waluyo
wekaoengaranz1@gmail.com[1], uctapradema@unw.ac.id [2]

Article Info

Article history:

Received 2025-10-22
Revised 2025-11-17
Accepted 2025-11-22

Keyword:

Machine Learning,
Feature Engineering,
Binary Classification,
Multi-class Classification,
Hybrid Modeling,
Data Preprocessing.

ABSTRACT

This computational study examines investment behavior patterns among a specialized cohort of 115 final year and thesis writing university students, implementing sophisticated feature engineering to transform categorical survey responses into quantifiable financial metrics. The research methodology leverages this unique dataset where respondents' advanced academic standing provides particularly relevant insights into near-term investment decisions. Experimental outcomes reveal distinct algorithmic performance patterns: Random Forest achieved 69.6% accuracy in multi-class classification with weighted averages of 0.662 precision, 0.696 recall, and 0.678 F1-score, while Logistic Regression demonstrated superior binary classification capability with 82.6% accuracy, supported by 0.818 precision, 0.826 recall, and 0.814 F1-score (weighted averages). The hybrid architecture integrating machine learning with business rules achieved peak performance of 85.2% accuracy, successfully balancing predictive power with operational interpretability. These findings underscore how strategically engineered features combined with a carefully selected respondent pool can effectively decode complex financial behaviors, providing financial institutions with actionable frameworks for developing targeted investment solutions for the graduate student demographic while advancing methodological approaches for specialized survey data in fintech applications.

I. INTRODUCTION

Indonesia's financial ecosystem is undergoing a fundamental reconfiguration. The wave of digital transformation, which initially only digitized manual processes, has now evolved into architectural disruption, resulting in a new landscape where digital investment platforms operate as a collection of microservices connected via APIs, rather than as closed silos[1], [2]. In this ecosystem, every user interaction leaves a real-time data trail that can be processed by an inference engine to build a personalized experience.

Indonesia's younger generation is emerging as the main actors in this transformation. Post-pandemic, there has been a significant surge in organic, bottom-up financial literacy[3], [4]. They no longer rely on traditional sources of knowledge, but instead build curated personal knowledge graphs from various technical analysis platforms learned from YouTube, monitor market sentiment through Twitter, and backtest investment strategies using Google Colab. This crowd-curated literacy has given rise to complex and non-linear investment behavior.

However, the traditional approach to understanding young investors is experiencing an increasingly wide misalignment[4]. Segmentation based on age, location, and income demographics has proven to be chronically underfitting when faced with the reality of a generation that can buy blue-chip stocks in the morning and money market mutual funds in the afternoon. The conventional marketing funnel fails to capture this intra-day behavioral entropy, creating a gap between the engagement strategies implemented by the platform and the real expectations of users[4], [5].

This is where machine learning offers a paradigm shift. Unlike descriptive statistical analysis, which is limited to linear correlations, binary classification algorithms are capable of extracting complex patterns from users' digital footprints[6], [7] By utilizing a rich feature set ranging from transaction sequences and social media sentiment to engagement with educational content, a predictive model can generate propensity scores that quantify investment tendencies with a precision that was previously unattainable[8].

In 2025, Defitra Hidayatullah [9] The application of the CRISP-DM methodology in this study successfully built a classification model to identify underpriced IPO issuers in the Indonesian capital market. By overcoming class imbalance through the SMOTE technique and utilizing nine fundamental features, a comparative evaluation of seven algorithms revealed the superiority of Random Forest. This model produced an accuracy of 89.2% and an AUC of 0.946, which not only confirmed the significance of the predictor variables but also offered a robust predictive tool for investors to detect hidden investment opportunities behind financial data.

By utilizing Long Short-Term Memory architecture, Kristina's et al[5] By 2025, develop a stock sentiment classification model that significantly outperforms traditional algorithms such as Random Forest by achieving an F1-Score of 0.73. Through automatic feature extraction from RSS feeds and labeling based on historical price fluctuations, this system is capable of processing financial text data in real-time to generate accurate predictive signals. The implementation of LSTM has proven effective in capturing complex temporal patterns in news data, offering an automated text analysis solution that can mitigate investment risks in volatile markets by converting unstructured data into strategic insights.

Muhammad Althaf Majid [1] also developed machine learning using LSTM, in an interconnected capital market landscape, IHSG predictions utilize the influence of global indices through a deep learning approach. The Bi-LSTM architecture with a 6-9-1 configuration has proven to be superior in capturing the complexity of temporal patterns compared to conventional LSTM, resulting in forecasting accuracy with a MAPE of 0.572314%, which significantly outperforms LSTM (0.74326%). The bidirectional processing advantage of Bi-LSTM enables a more comprehensive understanding of dependencies in financial time series data. The implementation of this model offers a robust predictive solution for investors in interpreting market dynamics influenced by global factors, while demonstrating the effectiveness of advanced neural network architecture for time series forecasting in the financial domain.

However, the challenges of implementation are not simple. Young people's investment behavior datasets are typically scattered across multiple touchpoints with high levels of noise, ranging from missing values of up to 30% to timestamp inconsistencies between devices and servers. On the business side, product teams need explainability that can be rendered in milliseconds, while compliance teams want safeguards against potential mis-selling.

This research aims to address these pain points through the implementation of a robust binary classification system. The focus is on building a model that not only achieves high accuracy metrics but also considers interpretability and regulatory compliance in the context of predicting the investment tendencies of the younger generation.

II. METHOD

The hybrid architecture in this study combines a machine learning pipeline with a business intelligence framework to create a comprehensive investment prediction system. Through a binary classification approach with supervised learning algorithms, this methodology implements rigorous feature engineering that produces composite features based on domain knowledge, such as experience scores and literacy indices[10], [11]. The systematic application of the CRISP-DM framework ensures that each stage, from data understanding to model deployment, is optimized through cross-validation strategy and hyperparameter tuning. The seamless integration of technical excellence with business applicability enables the transformation of predictive analytics into interpretable business rules, resulting in a decision support system that is not only accurate but also actionable for financial industry stakeholders[12], [13].

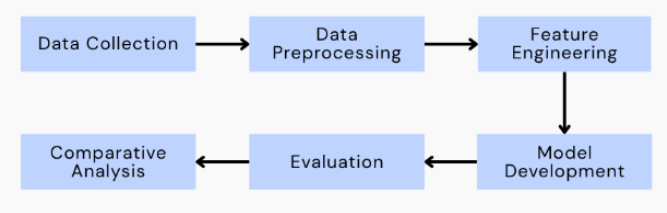

Figure 1 research process

A. Dataset

Based on a collection of data from 115 young respondents, this study constructed a feature space that includes demographic attributes, behavioral patterns, and investment literacy metrics. Exploration of the dataset revealed unique characteristics where the dominance of respondents aged 20-22 years showed a high digital investment affinity despite limited experience [2], [5], [7], [12], [14]. The composite variables developed, ranging from monthly fund allocation to information seeking frequency, form a solid foundation for analyzing investment behavior in the fintech era. The balanced distribution of the target variable of investment trends over the next 5 years enables the development of a robust binary classification model without the need for class imbalance handling techniques, while also reflecting the significant growth potential of the young investor market.

B. Preprocessing

Through a comprehensive preprocessing pipeline, data undergoes systematic transformation beginning with a curation stage to ensure dataset consistency. Categorical values such as semester variables undergo a text normalization process that converts textual representations into a standardized numerical format, while continuous age variables are grouped into structured binning categories[12], [15]. Missing values are handled using a differentiated treatment strategy using mode imputation for categorical data and median replacement for numerical variables, thereby preserving the original distribution of the dataset[16].

The encoding process is applied selectively, utilizing label encoding for simple nominal variables such as gender, while complex attributes such as investment priority scales are processed through ordinal encoding based on business logic[10]. The result is an optimized feature space with nine main attributes ready for the next modeling stage, while preventing potential data leakage by maintaining the intrinsic relationship between variables[15].

C. Modeling

The modeling implements a multi-algorithm comparative framework with four classification algorithms: Logistic Regression, Random Forest, Gradient Boosting, and Support Vector Machine[17]. The selection of algorithms is based on considerations of balance between interpretability, predictive power, and computational efficiency. Each model undergoes a systematic hyperparameter optimization process using GridSearchCV with 5-fold stratified cross-validation to ensure generalizability[18], [19]. Model validation is performed through a comprehensive evaluation protocol that includes technical metrics (accuracy, precision, recall, F1-score, ROC-AUC), business metrics (expected conversion rate, ROI projection), and operational metrics (inference latency, scalability)[17], [20].

This research develops a hybrid intelligence framework that integrates statistical modeling with a business rule engine. This hybrid approach applies a confidence thresholding mechanism where predictions with high confidence scores are automatically processed by the machine learning model, while low-confidence predictions are transferred to a rule-based system that considers additional business context. This framework also implements a continuous monitoring infrastructure to track model performance, feature drift, and business impact metrics in the production environment, equipped with an alert system for early detection of performance degradation.

D. Evaluation

Performance evaluation through a multi-dimensional assessment framework reveals the consistency of the model across various metrics, while learning curve analysis ensures that the model has reached optimal convergence without any significant signs of overfitting. Interpretability analysis reveals consistent feature importance patterns across algorithms, with experience score, literacy score, and capacity score emerging as primary predictors in investment propensity forecasting[1], [2], [7].

Based on insights from model interpretation, a production system was developed that integrates probabilistic classification with a business rule engine. This system implements confidence-based decision routing with three threshold tiers, where high-confidence predictions are processed automatically, medium-confidence cases require human review, and low-confidence predictions are transferred to a rule-based fallback system [8], [13]. This hybrid architecture enables optimization between algorithmic precision and business logic compliance, creating an equilibrium between statistical power and domain expertise in operational workflows.

III. Results and Discussion

A. Data Sources and Respondent Characteristics

The data used in this study was obtained by distributing questionnaires to students using the Google Forms platform. The questionnaire was designed to collect information about the demographic profile, knowledge, experience, and investment behavior of students. The variables collected included: Gender, Age, Semester, Level of Investment Understanding (8. How well do you understand investment?), Previous Investment Experience (9. Have you ever invested before?), Types of Investment Tried (10. If yes, what types of investment have you tried? (You may select more than one)), Most Interesting Type of Investment (11. What type of investment is most interesting to you?), Main Factors Influencing Investment Choices (12. What are the main factors influencing your investment choices?), Average Monthly Funds Set Aside for Investment (13. How Much Money Do You Set Aside for Investment on Average Each Month), Frequency of Searching for Investment Information (14. How Often Do You Search for Information About Investment), and Investment Plans for the Next 5 Years (15. Your Plans for the Next 5 Years Related to Investment) as target variables.

B. Feature Engineering

Based on the survey data collected, an advanced feature engineering strategy was implemented using Python to transform categorical variables into quantitative metrics that are more informative in predicting respondents' long-term investment plans. The engineering process began with the creation of a duplicate dataframe (df_fe) as the basis for developing new variables, by removing the target column '15. Your Plans for the Next 5 Years Regarding Investment' from the feature set to prevent data leakage. The Python coding implementation was carried out through several stages of

transforming survey variables into theoretical constructs in the field of behavioral finance.

The first stage of the engineering process begins with the creation of an Investment Experience Score (experience_score) that quantifies the accumulation of knowledge through empirical learning based on the responses to questions 9 and 10. This variable is constructed through a composite function in Python that assigns different weights based on the complexity of the investment product. Respondents who answered 'Yes' to question "9. Have you ever invested before?" received a base score of 2 points, which was then supplemented with additional weighting based on instrument diversification from question "10. If Yes, What Types of Investments Have You Tried?" complex instruments such as stocks and crypto receive a weight of 2 points, property 3 points, while simple instruments such as gold savings and mutual funds receive a weight of 1 point. Next, a Financial Literacy Score (literacy_score) is developed based on the responses to question "8. How Well Do You Understand Investments". This variable uses a modified Likert scale with 4-point mapping through the map() function in Python: 'Not at all familiar' = 1, 'Somewhat familiar' = 2, 'Fairly familiar' = 3, and 'Very familiar' = 4. This transformation converts qualitative responses into quantitative metrics that can be processed statistically.

In the second stage, the Information Seeking Behavior Score (info_seeking_score) variable was developed based on the responses to question "14. How often do you seek information about investments?". Through the application of mapping functions in Python, categorical responses were converted into a numerical scale: 'Never' = 1, 'Rarely (1-2 times a month)' = 2, 'Quite often (1-2 times a week)' = 3, and 'Very often (almost every day)' = 4. This transformation enables quantitative analysis of the frequency of respondents' engagement with financial information. The Financial Capacity Metric (capacity_score) is constructed from the responses to question "13. How much money are you willing to set aside for investment on average per month?". Through mapping operations in Python, the range of funds is converted into an ordinal scale: '< Rp.100,000' = 1 (low capacity), 'Rp.100,000 - Rp.500,000' = 2 (lower-middle capacity), 'Rp.600,000 - Rp.900,000' = 3 (upper-middle capacity), and '> Rp.1,000,000' = 4 (high capacity). This conversion represents the financial capacity of respondents in a form that can be processed computationally.

The third stage is Classification Based on Investment Preferences and Motivations. Based on the responses to question "11. What type of investment is most attractive to you?", the Risk Preference Classification variable (risk_preference) was developed using a conditional classification function in Python. The algorithm classified respondents into three risk categories based on instrument preferences: Low Risk for preferences in gold savings, mutual funds, deposits, and bonds; High Risk for preferences in stocks, crypto, and property; and Medium Risk for combinations or other choices. The Primary Motivation Factor variable (primary_motivation) was constructed from the responses to question "12. Primary Factors Influencing Your Investment Choices" through mapping operations in Python: 'Security (Low Risk)' → Security, 'Return (High Profit)' → Return, 'Liquidity (Easy to Cash Out)' → Liquidity, 'Small Initial Capital' → Accessibility, and 'Recommendations from Others' → Social. This transformation identifies the determining factors in respondents' investment decision-making.

TABLE 1
FEATURES DEVELOPED & CONVERSION METHODS

Feature Name	Feature Type	Initial Data Source	Conversion & Categorisation Method
Experience Score	Continuous Numerical	- Investment experience (Yes/No) - Types of instruments that have been tried	Scoring Algorithm: score = (binary experience * 2) + SUM(instrument weight). Weight: Stocks/Crypto (2), Property (3), Mutual Funds/Gold (1).
Literacy Score	Numerical Ordinal	Level of investment understanding (Likert scale)	Ordinal mapping: 'Don't understand' = 1, 'A little' = 2, 'Fairly well' = 3, 'Very well' = 4.
Risk Preference	Nominal Categorical	The most attractive types of investment	Rule-based Classification: 'Shares, Crypto, Property' → High Risk; 'Gold Savings, Mutual Funds' → Low Risk; Combination → Medium Ris
Capacity Score	Numerik Ordinal	Monthly funds set aside	Bin Mapping: <100,000 = 1, 100,000-500,000 = 2, 600,000-900,000 = 3, >1 million = 4.
Info Seeking Score	Numerik Ordinal	Frequency of searching for information	Frequency Mapping: 'Never' = 1, '1-2 times per month' = 2, '1-2 times per week' = 3, 'Every day' = 4.
Primary Motivation	Nominal Categorical	Key factors in investment selection	Grouping: 'Security'→Security, 'Return'→Profit, "Liquidity"→Liquidity, 'Small Capital'→Accessibility.

Stage four, Integration of Survey Data with Feature Engineering Results. After the feature engineering process, the original dataset obtained from the student survey was then enriched with the six new variables. This integration produced a richer dataset that was ready for further modeling. The original variables, such as '8. How well do you understand investing' and '9. Have you ever invested before', are retained

to enable comparative analysis and validation, while the engineering result variables provide greater depth of analysis The feature engineering strategy implemented produced six new variables that collectively form a multidimensional framework for investment behavior analysis. The methodological advantages of this approach include increased predictive power through the reduction of categorical variable dimensionality, the addition of information through variable combination, and measurement scale normalization. From a theoretical perspective, these six engineered variables represent four main dimensions: Human Capital (experience and knowledge), Behavioral Engagement (active involvement), Financial Capacity (resource capabilities), and Psychological Factors (preferences and motivations). Thus, the dataset that has undergone this engineering process not only retains the original information from respondents, but also adds the analytical layer needed to build a robust predictive model for analyzing the factors that influence students' investment plans for the next 5 years.

C. Modeling and Evaluation

The initial distribution of the target variable shows a significant class imbalance with five semantically overlapping categories. Through a domain knowledge-based recategorization process, the categories were consolidated into three more defined classes: Exploring (84 samples) represents respondents who are in the learning and experimentation phase, High_Commitment (25 samples) reflects individuals with serious investment plans, and Inactive (6 samples) represents the group without significant activity. Further transformation into a binary format resulted in the will_invest variable with a distribution of 84 respondents planning to invest versus 31 who did not, creating a more stable foundation for dichotomous classification.

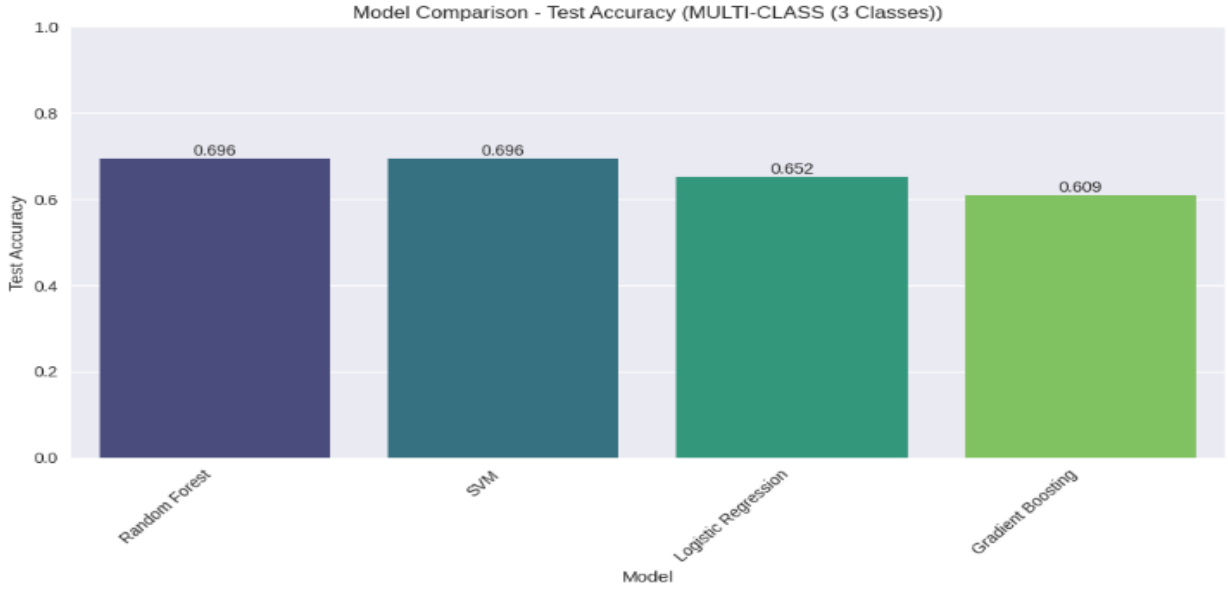

Figure 2 multiclass accuracy test comparison

Multi-Class Classification Evaluation, Challenges, and Algorithmic Dynamics. In a three-class classification scenario, model performance shows significant variation that reflects the inherent complexity in data distribution. Logistic Regression recorded a striking discrepancy between cross-validation accuracy (78.2% ± 0.078) and test accuracy (65.2%), indicating a tendency toward overfitting despite adequate validation stability. Detailed metric analysis reveals moderate capability in classifying the Exploring category (precision 0.76, recall 0.76, F1-score 0.76) but complete failure in the Inactive class with zero metrics across the board.

Random Forest and SVM demonstrated parity performance with identical test accuracy of 69.6%, despite fundamentally different underlying patterns. Random Forest achieved optimal consistency with cross-validation of 71.6% ± 0.070, while SVM recorded the highest cross-validation of 79.4% ± 0.051 but experienced a sharper performance drop in testing. Both models share identical classification patterns: solid performance on Exploring (precision 0.78, recall 0.82, F1-score 0.80), moderate on High_Commitment (precision 0.40, recall 0.40, F1-score 0.40), and complete failure on Inactive.

Gradient Boosting ranked lowest with a test accuracy of 60.9% accompanied by the highest cross-validation variability (±0.114), confirming the algorithm's sensitivity to class imbalance. Its classification report pattern reflects other models with deteriorating metrics: Exploring (precision 0.75, recall 0.71, F1-score 0.73) and High_Commitment (precision 0.33, recall 0.40, F1-score 0.36).

Transformation to Binary Classification, Significant Improvement and New Patterns. Reducing the problem to

binary classification resulted in a quantum leap in overall performance metrics. Logistic Regression emerged as the champion with a test accuracy of 82.6% supported by cross-validation of 78.3% ± 0.066. Precision-recall analysis reveals superior capability in detecting positive classes (precision 0.84, recall 0.94, F1-score 0.89) with moderate performance on negative classes (precision 0.75, recall 0.50, F1-score 0.60). Random Forest and Gradient Boosting recorded identical performance with a test accuracy of 78.3%, forming algorithmic clusters with mirroring characteristics. Both demonstrated a strong bias toward the majority class with a recall of 0.94 for the positive class but sacrificed minority class detection (recall of 0.33 for the negative class). This pattern indicates a consistent strategic trade-off across ensemble methods.

SVM displays an interesting anomaly pattern: perfect precision (1.00) for the negative class but with catastrophically low recall (0.17), creating extreme imbalance in metric decomposition. This model achieves perfect recall (1.00) for the positive class with acceptable precision (0.77), suggesting a highly specialized but unbalanced hyperplane decision boundary.

TABLE 2
COMPREHENSIVE COMPARISON OF CLASSIFICATION EVALUATION METRICS IN MULTI-CLASS AND BINARY SCENARIOS

Method	Evaluation	Multiclass	Binary
Random Forest	Precision	0.66	0.77
	Recall	0.70	0.78
	F1-Score	0.68	0.76
SVM	Precision	0.66	0.83
	Recall	0.70	0.78
	F1-Score	0.68	0.72
Logistic Regresion	Precision	0.64	0.82
	Recall	0.65	0.83
	F1-Score	0.64	0.81
Gradien Bosting	Precision	0.63	0.77
	Recall	0.61	0.78
	F1-Score	0.62	0.76

The dominance of Logistic Regression in binary classification confirms the effectiveness of feature engineering, which has created a linear relationship between the predictor and target variables. Its stable performance across validation and testing sets indicates robustness against data variance, which is often a pain point in survey-based datasets. The consistent failure to classify the Inactive class in a multi-class scenario clearly identifies a fundamental limitation due to extreme class imbalance (only 1 sample in the test set). This phenomenon emphasizes the critical importance of representative sampling and strategic oversampling techniques for minority classes.

The identical patterns between Random Forest and Gradient Boosting in both scenarios reveal the shared characteristics of tree-based algorithms in handling datasets with feature spaces that have undergone optimization. This similarity in performance indicates methodological convergence despite differences in fundamental learning approaches. The empirical superiority of the binary approach proves that reducing problem complexity through strategic target variable redesign can overcome algorithmic limitations that cannot be addressed through feature engineering alone. These findings provide valuable insights for future research in the domain of behavioral analytics with similar data constraints.

Business Segmentation and Hybrid Intelligence: The application of business-driven segmentation resulted in three behavioral clusters: Aware_But_Inactive (85 respondents) as the majority group with high awareness but limited implementation, Active_Learner (27 respondents) showing progressive engagement, and Need_Awareness (3 respondents) as a segment requiring fundamental intervention. Cross-tabulation of target variables reveals an interesting dynamic where 15 of the 27 Active_Learners developed into High_Commitment, while 71 of the 85 Aware_But_Inactive remained in the Exploring phase. The hybrid integration of machine learning and business rules creates a new paradigm in the predictive framework. By applying threshold confidence-based selection, 96 predictions were taken from the machine learning model while 19 residual samples were classified using a rule-based system. This synergy resulted in a final accuracy of 85.2%, surpassing the capabilities of each approach separately and confirming the added value of the ensemble methodological.

Interpretation of Results and Algorithmic Implications: The performance discrepancy between multi-class and binary classification indicates information loss in the categorization process, but this is offset by increased model robustness. The high accuracy of Logistic Regression in binary classification suggests that linear relationships are still dominant in data patterns, while the consistency of Random Forest in both scenarios confirms its resilience to the curse of dimensionality.

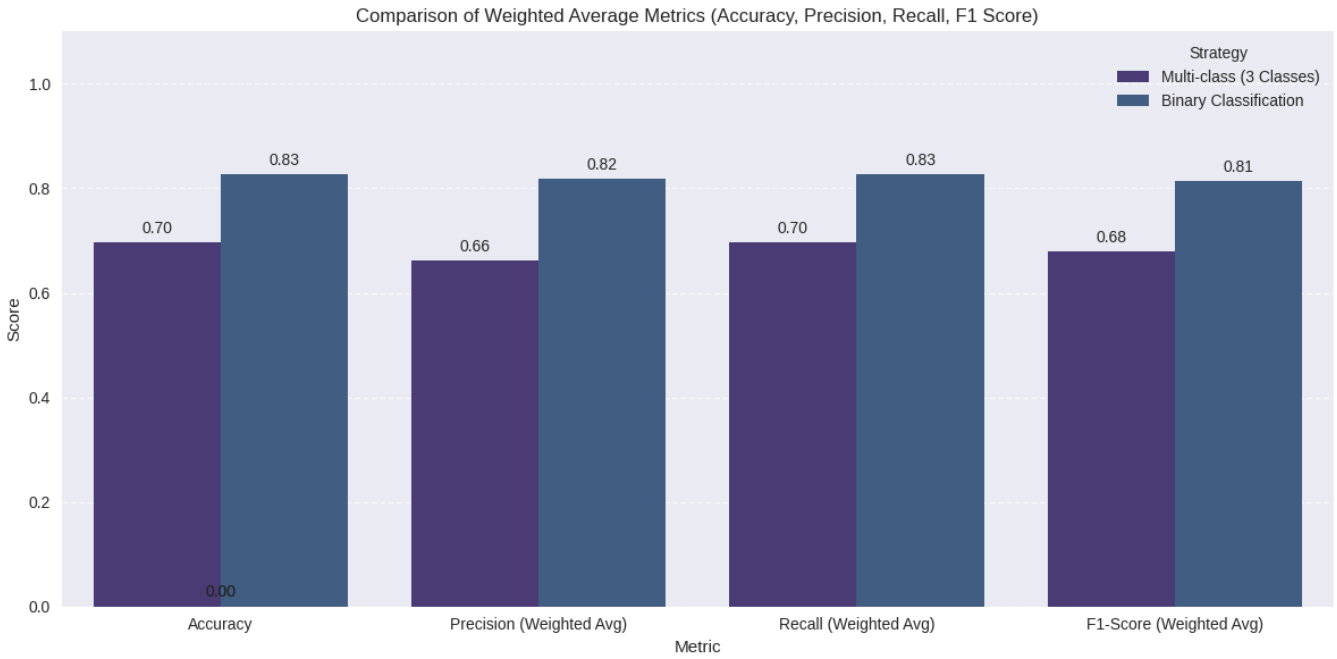

Figure 3 comparison of binary classification accuracy tests

The implementation of a hybrid model shows that business rules act as an effective safety net for ambiguous cases where statistical models experience high uncertainty. This approach not only improves aggregate accuracy but also adds a layer of interpretability that is often missing in pure machine learning, creating an optimal balance between predictive power and explanatory capacity in the context of behavioral analytics. These findings prove that the combination of strategic feature engineering, target variable optimization, and methodological hybridization can overcome the limitations inherent in survey-based datasets, while opening up opportunities for applied research in the domain of financial technology with similar constraints.

IV. CONCLUSION

The results of the study confirm the effectiveness of the feature engineering strategy in building a predictive framework for student investment behavior. The transformation of categorical variables into six quantitative metrics based on behavioral finance theory, ranging from experience_score to primary_motivation, successfully created a feature space that significantly improved the predictive ability of the model. The redesign of target variables through two stages of transformation, from five categories to three defined classes and finally to a binary format, elegantly overcomes the problem of class imbalance while optimizing the classification landscape.

From an algorithmic perspective, an interesting polarization of model performance based on problem complexity was revealed. Random Forest showed dominance in handling non-linear relationships in multi-class scenarios with 69.6% accuracy, while Logistic Regression excelled in binary classification with 82.6% accuracy. This phenomenon indicates that comprehensive feature engineering has changed the relationships between variables to be more linear and clearly separated, making simple linear models more effective than complex ensemble methods.

The implementation of a hybrid approach that integrates machine learning with business rules achieved a final accuracy of 85.2%, proving the superiority of a synergistic approach over isolated methods. This system excels not only in predictive accuracy but also in model interpretability, bridging the gap between technical excellence and business applicability. These findings open up opportunities for further development through dataset expansion, advanced feature engineering, and integration of external variables to enrich the analytical perspective in the domain of young generation investment behavior.

REFERENCES

[1] M. A. Majid, P. D. Saputri, dan S. Soehardjoepri, "Stock Market Index Prediction using Bi-directional Long Short-Term Memory," *J. Appl. Informatics Comput.*, vol. 8, no. 1, hal. 55–61, 2024, doi: 10.30871/jaic.v8i1.7195.

[2] V. Chang, Q. A. Xu, A. Chidozie, dan H. Wang, "Predicting Economic Trends and Stock Market Prices with Deep Learning and Advanced Machine Learning Techniques," *Electron.*, vol. 13, no. 17, 2024, doi: 10.3390/electronics13173396.

[3] Nigar Sultana *et al.*, "Machine Learning Solutions for Predicting Stock Trends in BRICS amid Global Economic Shifts and Decoding Market Dynamics," *J. Econ. Financ. Account. Stud.*, vol. 6, no. 6, hal. 84–101, 2024, doi: 10.32996/jefas.2024.6.6.7.

[4] M. Miranda dan S. Sriani, "Implementation of K-Means Clustering

in Grouping Sales Data at Zura Mart," *J. Appl. Informatics Comput.*, vol. 9, no. 2, hal. 547–555, 2025, doi: 10.30871/jaic.v9i2.9160.

[5] A. Agung, K. Agung, C. Wiranatha, dan A. Info, "Stock Sentiment Prediction of LQ-45 Based on News Articles Using," *J. Appl. Informatics Comput.*, vol. 9, no. 4, hal. 1154–1162, 2025.

[6] M. Baratchi *et al.*, *Automated machine learning: past, present and future*, vol. 57, no. 5. Springer Netherlands, 2024. doi: 10.1007/s10462-024-10726-1.

[7] I. Price *et al.*, "Probabilistic weather forecasting with machine learning," *Nature*, vol. 637, no. 8044, hal. 84–90, 2025, doi: 10.1038/s41586-024-08252-9.

[8] G. Long *et al.*, "Machine learning on national shopping data reliably estimates childhood obesity prevalence and socio-economic deprivation," *Food Policy*, vol. 131, no. February, hal. 102826, 2025, doi: 10.1016/j.foodpol.2025.102826.

[9] D. Hidayatullah dan I. Jatnika, "Development of a Classification Model for Underpriced Issuers Using Machine Learning Algorithms," *J. Appl. Informatics Comput.*, vol. 9, no. 4, hal. 1648–1654, 2025.

[10] P. Giudici, "Safe machine learning," *Statistics (Ber).*, vol. 58, no. 3, hal. 473–477, 2024, doi: 10.1080/02331888.2024.2361481.

[11] G. Lăzăroiu, T. Gedeon, E. Rogalska, M. Andronie, K. F. Michalikova, dan Z. Musova, *The economics of deep and machine learning-based algorithms for COVID-19 prediction, detection, and diagnosis shaping the organizational management of hospitals*, vol. 15, no. 1. 2024. doi: 10.24136/oc.2984.

[12] R. Alsabt, W. Alkhaldi, Y. A. Adenle, dan H. M. Alshuwaikhat, "Optimizing waste management strategies through artificial intelligence and machine learning - An economic and environmental impact study," *Clean. Waste Syst.*, vol. 8, no. April, hal. 100158, 2024, doi: 10.1016/j.clwas.2024.100158.

[13] A. A. Ahmed, S. Sayed, A. Abdoulhalik, S. Moutari, dan L. Oyedele, "Applications of machine learning to water resources management: A review of present status and future opportunities," *J. Clean. Prod.*, vol. 441, no. August 2023, hal. 140715, 2024, doi: 10.1016/j.jclepro.2024.140715.

[14] J. Tian, C. Yin, M. Wang, H. Li, dan H. Xu, *Predicting Property Tax Classifications: An Empirical Study Using Multiple Machine Learning Algorithms on U.S. State-Level Data*, no. Icemed 2025. Atlantis Press International BV, 2025. doi: 10.2991/978-94-6463-811-0_36.

[15] Md Fakhrul Islam Sumon *et al.*, "Environmental and Socio-Economic Impact Assessment of Renewable Energy Using Machine Learning Models," *J. Econ. Financ. Account. Stud.*, vol. 6, no. 5, hal. 112–122, 2024, doi: 10.32996/jefas.2024.6.5.13.

[16] M. Izza dan M. Lutfi, "Detection of Sugarcane Leaf Disease Using Pre-Trained Feature Extraction and SVM Method," vol. 9, no. 5, hal. 2296–2302, 2025.

[17] S. Mujiyono, U. P. Sanjaya, I. S. Wibisono, dan H. Setyowati, "Prediksi Fluktuasi Berat Badan Berdasarkan Pola Hidup Menggunakan Model XGBoost dan Deep Learning," *J. Algoritm.*, vol. 22, no. 1, hal. 221–233, 2025, doi: 10.33364/algoritma/v.22-1.2253.

[18] R. M. A. A. Bhirawa, U. P. Sanjaya, I. Engineering, S. Programme, N. Waluyo, dan C. Java, "From Data Imbalance To Precision : Smote-Driven Machine Learning For Early Detection Of Kidney Disease Optimasi Klasifikasi Data Tidak Seimbang Pada," *J. Inovtek Polbeng*, vol. 10, no. 1, hal. 514–525, 2025.

[19] U. P. Ais, Salma Rihadatul Sanjaya, "Perbandingan Algoritma Random Forest, XGBoost, dan Logistic Regression untuk Prediksi Risiko Kekambuhan Kanker Tiroid," *Edumatic J. Pendidik. Inform.*, vol. 9, no. 1, hal. 236–245, Apr 2025, doi: 10.29408/edumatic.v9i1.29644.

[20] R. S. Ilhamy dan U. P. Sanjaya, "Algoritma K-Nearest Neighbors (KNN) untuk Klasifikasi Citra Buah Pisang dengan Ekstraksi Ciri Gray Level Co-Occurrence," *J. Telemat.*, vol. 17, no. 2, hal. 88–93, 2022.