Journal of Computer Science

Original Research Paper

Development of Customer Predictive Model for Investment Using Ensemble Learning Technique

Thongchai Kaewkiriya and Kittipol Wisaeng

Mahasarakham Business School, Mahasarakham University, Mahasarakham, Thailand

Article history
Received: 28-12-2022
Revised: 23-03-2023
Accepted: 13-05-2023

Corresponding Author:
Kittipol Wisaeng
Mahasarakham Business School, Mahasarakham University, Mahasarakham, Thailand
Email: kittipol.w@acc.msu.ac.th

Abstract: Many new investors that want to purchase a fund might be disappointed with the return-on-investment value. This problem occurs because they do not know important factors that could affect the market value. Plus, some assets might not be suitable for the investor's investment style. To reduce the mentioned problems, many asset management companies research the systems that could find a suitable fund for the investor. These systems consider various aspects, such as the investor's background, financial status, and the investor's behavior. These systems typically employ a specific machine learning method to learn and predict which fund the model should recommend to customers. However, we have an assumption that the performance of the prediction could be leveraged if we applied more methods to a forecasting model. Therefore, this study aims to develop a customer predictive model for investment using voting ensemble learning techniques. The model is used for recommending suitable funds and suitable risks for investing based on the investor's profile and comparing performance between 5 algorithms and 2 preprocessing approaches. Preprocessing approaches are clustering by date range which has an average accuracy of 62.24% and k-means clustering which has an average accuracy of 69.21%. The prediction model of suitable fund risk and prediction model of fund category has an accuracy of 92.38%. We found that the neural network has the highest accuracy of 93.43%.

Keywords: Data Preparation, Ensemble Learning, Investment, Machine Learning

Introduction

Nowadays, there are many choices of funds and assets for investors to start an investment. Plus, with technology, they are easier to obtain and gain benefits from. However, in the real world, the investment has many factors to consider and some risks could affect the return on the investment. Furthermore, some types of assets require prior knowledge, which might not be appropriate for some investors to invest in. Novice investors or someone who is not familiar with the investment field might find it difficult to analyze these important factors. These issues could result in an unsatisfactory profit or even a significant loss of the asset. The stated problems have become a challenge for the asset management company to develop a system or model to help the investors' decision on which asset or fund to invest in.

To develop such a system, the researchers apply the concept of machine learning to create a predictive model for the fund's recommendation. The model will suggest the appropriate funds to invest in based on an investor's data, such as their financial status, type of saving, investor behavior, or risk management. Nevertheless, we have an assumption that each dataset is suitable for a different method of machine learning. Therefore, choosing only one machine learning approach and training with every dataset might be insufficient to achieve a high-performance funds recommendation model.

To prove our hypothesis, this study aims to develop a customer predictive model for investment using the voting ensemble learning technique for recommending suitable funds and suitable risks for investing based on an investor's profile. This also includes performance comparison of algorithms and preprocessing approaches before model training. In the testing phase, the baselines and our model are evaluated with the cross-validation method. We also divided the training and testing data into different ratios to observe the consequences that might affect the performance of the prediction.

Therefore, the main contributions of this study are listed as follows:

1) The voting ensemble learning technique helps improve the performance of the fund recommendation model and can evaluate the risk of investing effectively
2) The ratio of the training and testing data does not affect the performance of the prediction model. Hence, the researcher can train the model with the lowest number of training samples to reduce the processing time

Materials and Methods

Concept of Investment

Investment is the idea of using saving money to invest in something to receive more money in return e.g., deposit money with a bank to receive interest. However, the return of interest is too small and unattractive. Hence investors look for somewhere else to invest to get a high return (SangSoi, 2015) every investment comes with risk, the investment can be separated into three categories: (i) Investment for consumption, (ii) Investment in the business and (iii) Investment in securities (Sungkaew, 2001).

Data Mining

Data mining is a technique for discovering relationships and extracting useful features from large amounts of data. This technique helps many applications by providing useful information for decision-making in determining tasks. An example of the data mining process is the cross-industry standard process for data mining. It consists of six procedures: Business understanding, data understanding, data preparation, modeling evaluation, and deployment. These processes have been applied to our paper since they are commonly used in various data-minimizing-related searches (Berwind *et al*., 2016).

Clustering

K-means clustering is an algorithm that partitions data into determined groups by using the average distance between data points and the centroid to categorize the group. Data with similar properties and features will be clustered with this approach.

Data clustering can be used to preprocess data to reduce the size of the data or simply select or exclude clusters for further study (Ahmad and Dey, 2007) (MacKay, 2003).

Classification Algorithm

A classification algorithm is an algorithm that learns through patterns of data from the samples and then classifies data based on their properties or features. In this study, we study 5 Algorithms as follows.

K-Nearest Neighbor Algorithm

This algorithm compares the similarity between data points and will put the input data in the same class as the data point that is closest to Hulett *et al*. (2012); Piryonesi and El-Diraby (2020). The distance can be calculated by using a distant equation, e.g., the Euclidean distance (Hastie *et al*., 2009) as shown in formula 1, when *x*, *y* is the data coordinate and *a*, *b* is the target coordinate that we want to find the distance between them. K-NN can create an effective model even though the decisions are complex, but it can take a long time if there are a lot of attributes:

$$Dist((x,y),(a,b)) = \sqrt{(((x-a) \wedge 2+(y-b) \wedge 2))} \quad (1)$$

$$a+b=c \quad (2)$$

Naïve Bayes Algorithm

This algorithm is about the probability of something happening when its condition has been qualified, called "given" (Hastie *et al*., 2009; Murty and Devi, 2011). The goal is to find which assumption is the most likely based on prior knowledge. This can be explained by dividing the number of desired results by the number of every possibility. The possibility of event *A* happening based on event *B* is called conditional probability. This algorithm calculates the probability of each result which can be simplified as shown in formula 2, where $P(A|B)$ is the probability of event *B* to happen which is required for event *A* to happen or called *A* given *B*. $P(A)$ is the probability of event *A* happening. $P(B)$ is the probability of event *B* happening. $P(B|A)$ is the probability of *B* given *A*:

$$P(A|B) = P(B|A) * P(A) / P(B) \quad (3)$$

Decision Tree Algorithm

This algorithm creates prediction models based on conditional classification by analyzing data features. This can be used to explain which feature has a high influence on the model's decision-making. The model created by this algorithm is in the form of a tree that consists of (i) Node which is a data feature where the first node is the root node, (ii) Branch which is the criteria of said features and (iii) A leaf which is the classified class. The decision tree is created by applying the greedy approach and using a top-down recursive divide-and-conquer approach. This algorithm learns by dividing a dataset into smaller sets during

feature selection to reduce the number of variables that are needed to create the model by selecting the best or most important features for prediction. The feature selection can be divided into two categories, as follows.

Feature's Subset Selection

Select only a few subsets of features from all features. The selected subset should improve the accuracy of the model.

Feature Sorting

Calculating a score for each feature sorting the feature according to the score. The equation is shown in Eq. 3:

$$IG(X;Y)=H(Y)-H(Y|X) \quad (4)$$

where, $IG(X;Y)$ is the entropy score that is a value between 0 and 1, $H(Y)$ is the probability of Y entropy, $H(Y|X)$ is the probability of Y given X entropy, Y is the value of the feature which is classes of data from $\{Y_1, Y_2, \ldots, Y_n\}$ and value of X is another feature that is not class $\{X_1, X_2, \ldots, X_n\}$.

Rule Induction Algorithm

Rule induction algorithm is an algorithm that does not need a human to program manually, but it will analyze the data structure or rules set to classify data. It can create independent rules and does not need to be in a hierarchy form. This algorithm can find different patterns when compared to the decision tree and may create a better classification model (Freitas, 2002; Cohen, 1995).

Neural Network Algorithm

This is an algorithm that tries to replicate the structure of the brain of living beings, which can learn and adjust according to input based on learning rules (Anthony and Bartlett, 2017). This algorithm is also the first pattern recognition algorithm to outperform human proficiency in the (Faggiolani, 2011) competition. Neural networks consist of (i) An input layer that only accepts numerical values, (ii) An output layer that is the result of the learning, a hidden layer that calculates values for prediction, (iii) Neurons that are in each layer that have different functions, (iv) Weights which are calculated during the learning process, (v) Bias which is calculated to help with decision boundary for more accuracy, (vi) Summation function which summarizes data input and its weight and (vii) Transfer function which calculate how to send the result to another node.

Ensemble Learning Technique

It is a machine learning technique that enhances prediction performance by utilizing multiple classification models to find the result, which is a highly effective approach (Dietterich, 2000; Manish, 2012) and also practical algorithms for a specific prediction task (Hopfield, 1982). The ensemble algorithm that is used in this study is called vote ensemble, which is the training of the same dataset on multiple prediction models, then picking the answer by counting from the output of each model, which can be called voting.

Fund Recommendation Models

Starting from research on behavior study and customer needs, there is a model (Tanizaki *et al*., 2020) that uses customer needs to organize the shop, e.g., recruiting employees and food ordering. The study utilizes machine learning to learn from information, e.g., weather.

We can see that the amount of customer data and other variables can be overwhelming, so we need some ways to handle the data. The state-of-the-art solutions to solve the mentioned problems are data mining and machine learning (Zhang *et al*., 2018). For instance, using k-means clustering to cluster customers together based on their preferences and recommend a restaurant for the user. The model is evaluated by comparing it with a case study of TripAdvisor.com in the aspect of investment.

Another work by Sapaphan (2016) shows that it is appropriate to use a decision tree to create a prediction model with preprocessing.

There is also a study by He *et al*. (2014) that tries to predict customer attrition with a support vector machine and found that this method can be tuned to achieve higher accuracy.

For the development of a customer investing prediction model by using ensemble learning, we have compared and studied the y working process of each algorithm. There is a study (Jaiswal *et al*., 2020) that try to predict customer transaction by comparing deep learning, XGBoost, and logistic regression. They found that deep learning has the most accuracy followed XGBoost and linear regression respectively. Another study tries to try to predict liquidity rather o of mutual funds via ensemble learning which is a majority vote. The result of that work has good interpretability and is beneficial to every involved party. Tao *et al*. (2019) proposed an approach to classify mutual fund investment types on the data from Yahoo Finance which consists of 25,393 funds and 54 features, then comparing the performance of 4 algorithms i.e., k-nearest neighbors, neural network, XGBoost, and random forest. The said research concludes that XGBoost has the best performance and the neural network performed the worst.

From what was mentioned above, this study has two main study objectives: (1) How preprocessing affects the prediction accuracy and (2) Comparing the performance of each algorithm to find the best predictive model.

Problem Analysis and Research Dataset

To analyze the problem, data, factors, and product presentation from the asset management company, we study them through the following steps:

1. Study the data, problems, documents, and related research: This process is to summarize and study the trend of the developing prediction model. This also includes studies about fund or asset selling procedures, e.g., product presentation, product review, and closing the deal
2. Understanding the research variables: Collecting customer data that bought funds from a certain bank. This process consists of four important steps: (i) Validating the data to ensure the completeness of the samples, (ii) Considering and choosing the important factors that could affect the research objectives, (iii) Data cleaning and (iv) Transforming the data into the form that the algorithm requires. The dataset for the experiment in this study is from the anonymous asset company, which contains 19,577 customers that bought funds between January and July 2015 and needed to purchase at least 500 baht. The fund can be categorized into four categories, in Table 1

Then, we collect other variables of the customer population to be used as input and to develop a prediction model. The variables are shown in Table 2.

Framework

The illustration of the framework is depicted in Fig. 1 and the process of this framework is described as follows.

Data Handling

This process can be divided into two approaches, as follow:

1. Approach: Transforming the data by using a number to represent a range of data, which are divided equally, in Table 3
2. Approach: Transforming the input and output by using the k-means algorithm. This can be done by separating data into appropriate k-groups before using it to develop a prediction model

Prediction Model Development and Performance Comparing

After the preprocessing data in the previous step is complete, those data are fed into different machine learning techniques i.e., K-NN, Naïve Bayes, decision tree, rule induction, and neural network. The reason that we chose these algorithms are as follow:

1. K-NN and Naïve Bayes feature different approaches as an algorithm which is a lazy algorithm and Bayesian respectively, so we can compare the differences between them
2. All of them use the same format of the input. This helps reduce complexity and bias when performing data cleansing
3. These are basic algorithm that is the base of other algorithms e.g., a decision tree is the root of the random forest and a neural network is the base of deep learning (Kotsiantis *et al*., 2007)

The accuracy is evaluated by dividing data into testing data and learning data. We also test every model with the cross-validation technique and adjust parameters to find the parameter that has the highest impact on the model accuracy.

Development of Prediction Model Using Ensemble Learning Technique and Performance Comparison

Developing a prediction model using an ensemble-learning technique by voting and evaluating the model. The evaluation in this part is similar to the second part but combined with the data handling approaches from the first part to develop two models, e.g., a prediction model on fund category which recommends the suitable category, and a prediction model on fund risk which recommends the suitable fund risk to give more details.

Table 1: Some information on fund categories

Class	N	Average	Min	Max	S.D
LTF/RMF (LR)	2,992	39,911	500	900,000	69,195.33
Money Market (MM)	11,562	794,790	500	9,850,000	1,171,852.40
Mutual Fund (OE)	6,441	932,044	1,085	9,472,406	659,896.77
Team Fund (TF)	3,485	270,921	500	8,000,000	1,041,179.52

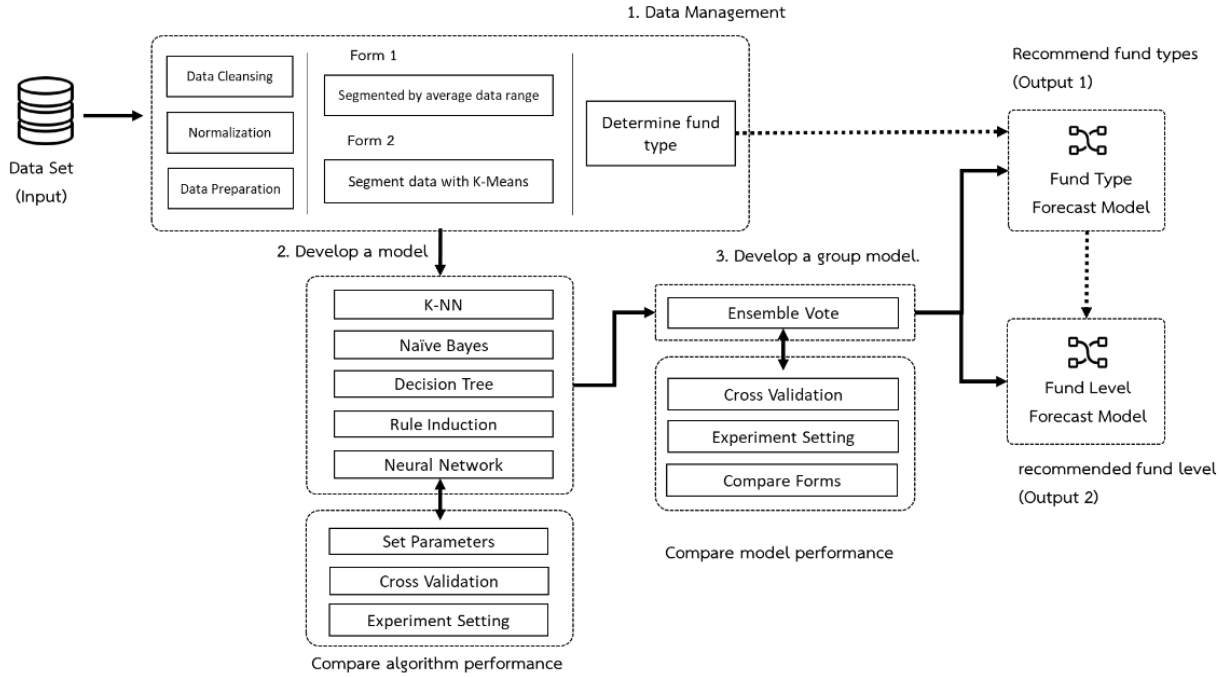

Fig. 1: The customer investment prediction model

Table 2: Overview of the input variable

Variable	Meaning
Branch_Provice	Location
Gender	Gender
Education	Education level
Occupation	Occupation
Income	Income
Saving A/C	Saving account
Fixed A/C	Fixed account
SPA A/C	Special fixed account

Table 3: Example of data transformation in the first approach

Class	Date range (fund price range)
LTF/RMF (LR) class label	200,000-1=”1”
	400,000-200,001=”2”
	600,000-400,001=”3”
	800,000-600,001=”4”
	1,000,000-800,001=”5”
	>1,000,000=”6”

Results

The details of the analysis according to the research goal are as follows.

The Details of Variables That are Used to Develop Prediction Model on Customer Investment

From this study, the variables that are necessary to develop a predicting model are shown in Tables 4-5.

The Result of Preprocessing Data with the First Approach for Prediction Model on Suitable Fund Risk

After collecting real variables and transforming them into LTF/RMF (LR), Mutual Fund (MF), Team Fund (TF), Special fixed deposit Account (SPA), and saving account (Saving A/C), we found that most of the data fall into the 1-200,000 range, which is 97.46, 38.05, 75.41, 44.50 and 38.95% of each variable, respectively, while Mutual Fund (MF) and fixed deposit account have the data in the range of over 1,000,000, which is 27.54 and 44.89%, in Fig. 2. LTF/RMF (LR) has no data that goes into more than the 1,000,000 range.

The Result of Preprocessing Data with a Second Approach for Prediction Model on Suitable Fund Risk

After finding the best *K* value for each fund, we can conclude that K = 5 is the best for LR and K = 4 is the best for MF, TF, and MM. The best value is determined by comparing it with different K-values.

The Result of Comparing Each Algorithm Performance

After determining and tuning parameters for the performance comparison, we evaluate each model with 5-fold cross-validation. The results of the two preprocessing approaches are as follows.

Table 4: The nominal variables that are used to develop a predicting model

Variables	Value	Count
Branch province	Bangkok and surrounding areas	8666
	North	2459
	Central	2893
	South	3427
	Northeast	2131
Gender	M	13411
	F	6165
Education	Lower than a bachelor's degree	11308
	Master's degree	3009
	Bachelor's degree	4860
	Higher than a master's degree	399
Occupation	Private employee	7726
	Government officer	1761
	Freelance	7894
	Personal business	2195
Income	20,001-40,000	8004
	60,001-80,000	6927
	1-20,000	2306
	40,001-60,000	1405
	> 100,000	291
	80,001-100,000	643

Table 5: The real variables that are used to develop a predicting model

Value	N	Max	Average
Fixed A/C	4664	4873236.58	1179798.620
SPA A/C	3986	4572914.77	503593.170
Saving A/C	19563	4992329.23	747393.160
Fixed A/C	4664	4873236.58	1179798.620

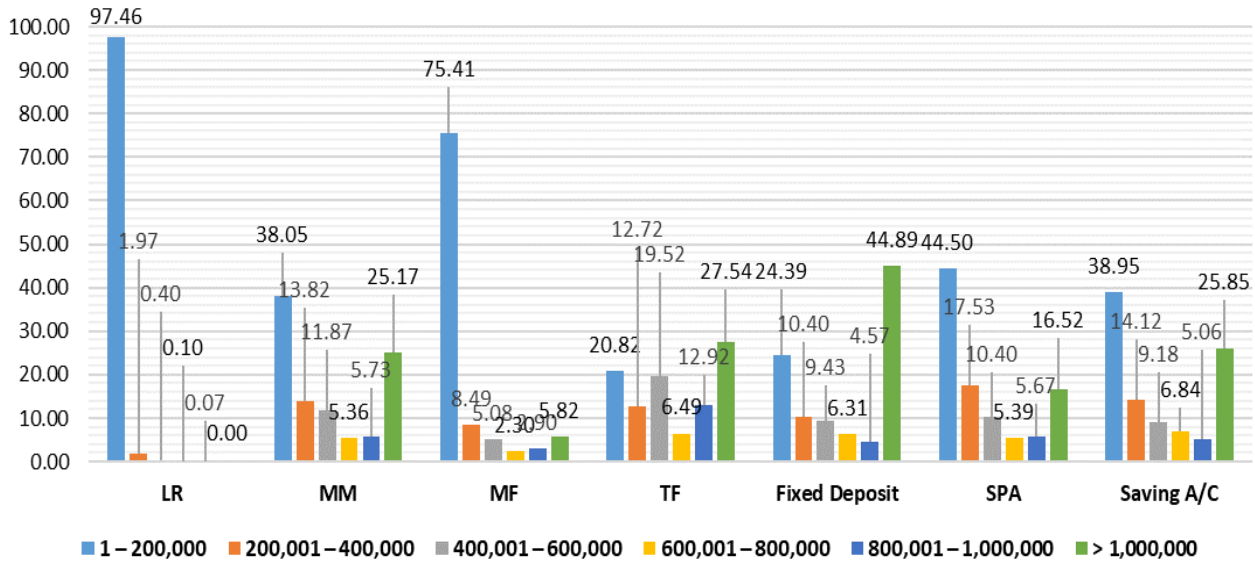

Fig. 2: The percentage of data after being categorized by range

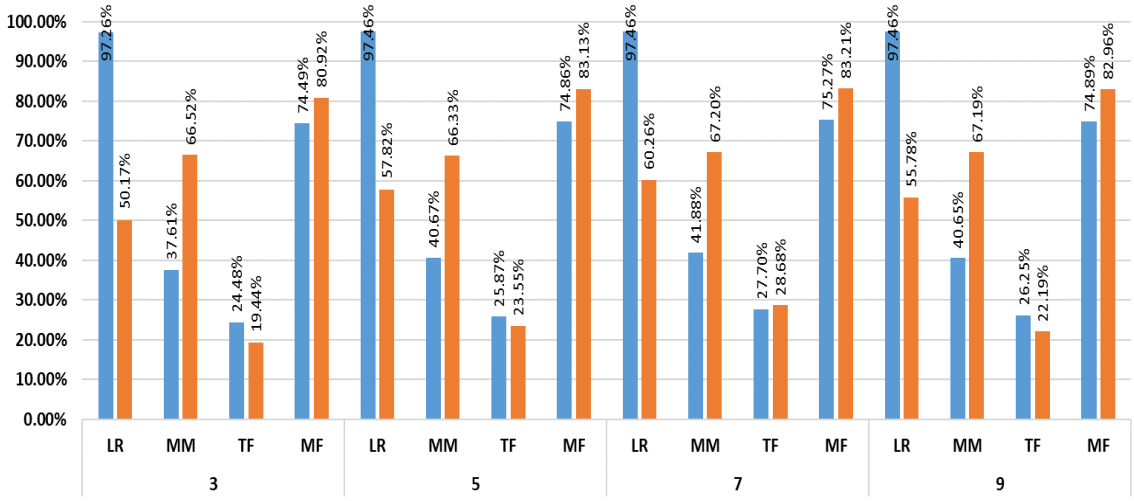

Fig. 3: The accuracy of a k-nearest neighbor when using different preprocess approaches

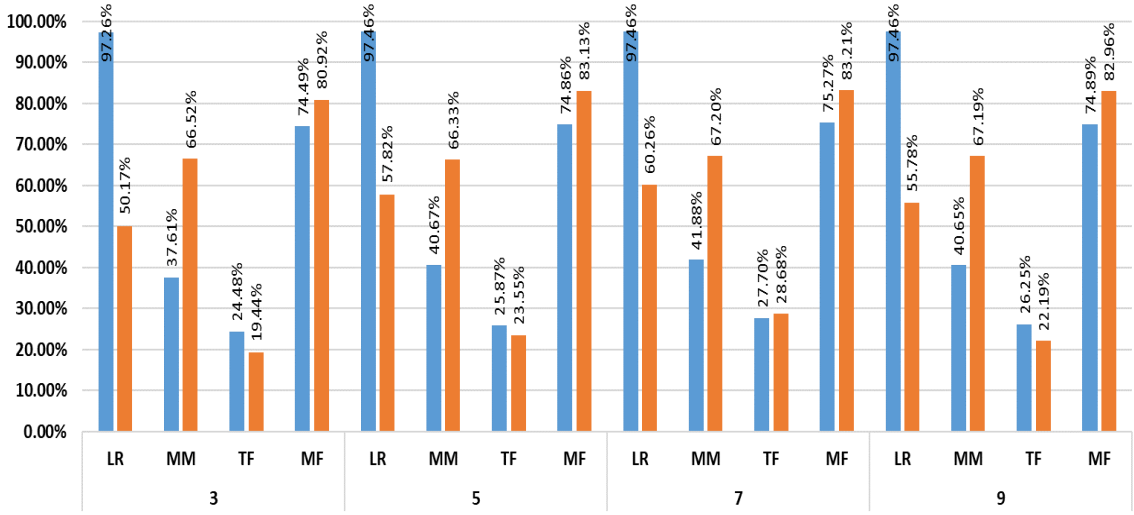

Fig. 4: The accuracy of the k-nearest neighbor when using different preprocess approaches

K-nearest neighbor algorithm: We found that different K-values do not have any effect on the model accuracy. The accuracy of this model when combined with preprocessing by k-means compared to preprocessing by range shows huge differences. e.g., in the LR category, k-means has 97.26% accuracy, and preprocessing by range has 50.17% accuracy. While performing under the TF category, k-means has 24.48% accuracy, and preprocessing by range has 19.44% accuracy. Data distribution may affect the small change in the prediction model in Fig. 3.

Naïve Bayes algorithm: There is no additional parameter. The accuracy of this algorithm is comparable to other algorithms, but the LR category has reduced accuracy while other categories have higher accuracy.

Decision tree algorithm: There is no change in accuracy when changing the criterion value, but when comparing preprocessing approaches, there is a huge shift in the performance. e.g., in the LR category, the gain ratio of preprocessing by range has higher accuracy than preprocessing by k-means at 97.36 and 69.95%, respectively, while preprocessing by k-means can perform better in other categories, in Fig. 4.

Rule induction algorithm: There is no change in performance when the criterion value is changed. We found that information gain between 2 preprocessing approaches has a different value, e.g., the preprocessing by range has 97.46% accuracy, while k-means has 69.62% accuracy under the LR category, while k-means outperform in other categories, in Fig. 5.

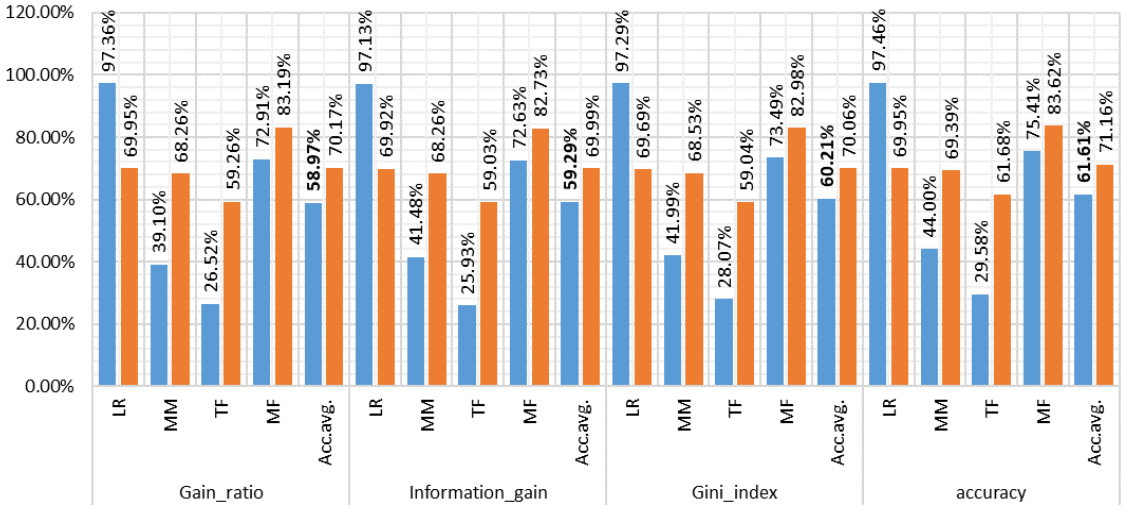

Fig. 5: The accuracy of the decision tree when using different preprocess approaches

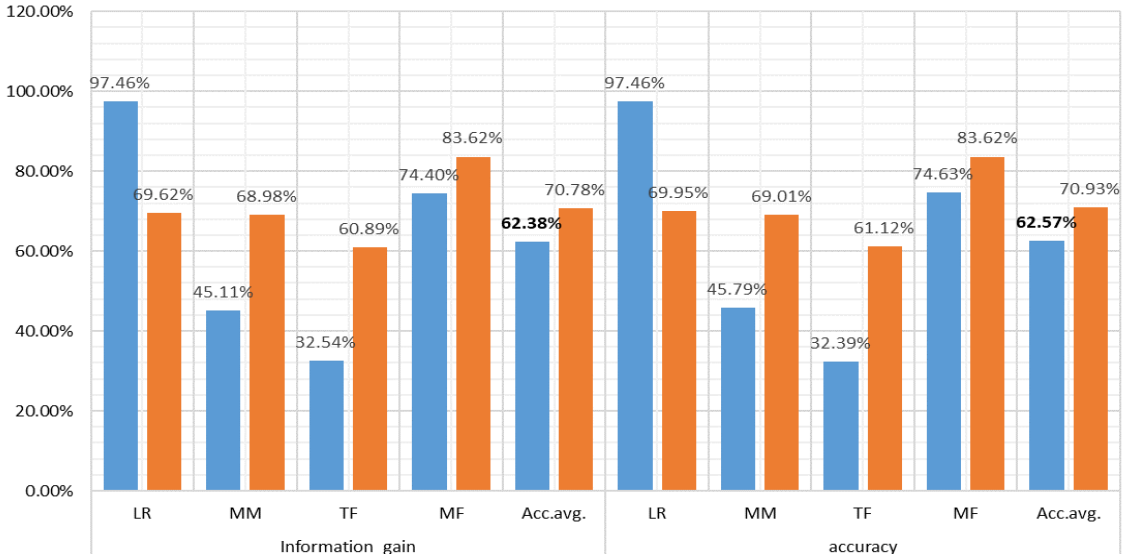

Fig. 6: The accuracy of rule induction when using different preprocess approaches

Neural network algorithm: There is no change in performance when the criterion value changes. We found that preprocessing approaches have a strong impact on the accuracy of the model, in Fig. 6.

After considering the average accuracy of each fund category and the preprocessing setting, the neural network algorithm has the highest accuracy (67.32%) followed by Naïve Bayes (67.09%), rule induction (66.77%), decision tree (66.39%) and k-nearest neighbor (60.58%). If we preprocess with k-means, the accuracy of each model is as follows: Neural network (71.18%), Naïve Bayes (63.08%), decision tree (71.16%), rule induction (70.93%) and k-nearest neighbor (59.84%), respectively, in Fig. 7. This means that the k-nearest neighbor algorithm is not suitable for this task. Ensemble learning can attain high accuracy, but it requires more processing time to process data and we found that the accuracy of another algorithm that is not k-nearest neighbor can attain comparable accuracy.

The Model Performance Comparison Based on the Data Setting Splitting Method

We evaluate each model by splitting the available data into six sets. Each set contains a different ratio of the training and testing data. They consist of 80-20, 70-30, 60-40, 50-50, 40-60, and 30-70, respectively. The result of this experiment is shown in Fig. 8. The K-NN has the lowest accuracy score, which is 57-59%. We also find that the different ratio of the training and test data does not affect the performance of the prediction model.

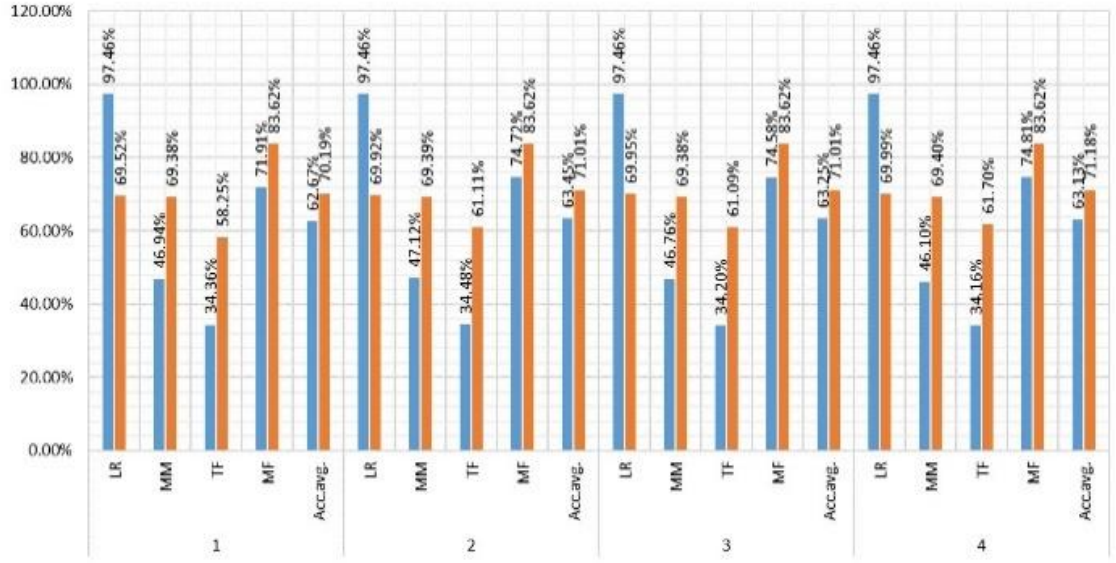

Fig. 7: The accuracy of the neural network when using different preprocess approaches

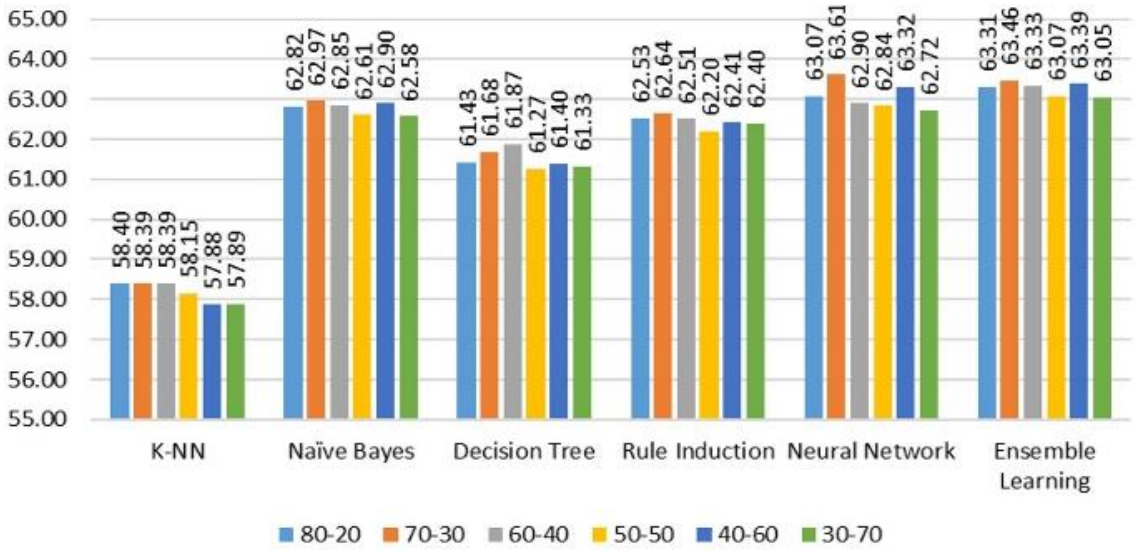

Fig. 8: The average accuracy of each algorithm represent in different ratios of training and testing data

This interesting discovery could be beneficial for the future of other related research. If the size of the training set does not tremendously affect the performance of the prediction model, we can reduce the training samples to reduce the complexity of the model and greatly reduce the computational time. The supporting research for this statement is reported by Rolim *et al*. (2021).

Conclusion

In the development of a customer investment predictive model by using the voting ensemble technique from five based algorithms, we found that the preprocessing approach has an impact on every algorithm. The average accuracy of data that is preprocessed by the first approach (representing data by its range) is 62.24% while preprocessing by k-means clustering has a 69.21% average accuracy. As a result, we can conclude the evaluation of our experiment based on the consideration factors as stated below:

1. Considering fund categories: LR category has the highest average accuracy at 97.42% when preprocessed with clustering data by range, while MM, TF, and MF have the highest accuracy at 68.95, 56.91 and 83.55% respectively when preprocessed with a k-means algorithm

2. Considering each algorithm: In the LR category, every algorithm shares the highest accuracy at 97.46%. In the MM category, the accuracy of each algorithm is as follows: Neural network at 69.40%, decision tree at 69.39%, Naïve Bayes at 69.33%, rule induction at 69.01%, and k-nearest neighbor at 67.20%. After combining it into an ensemble model, the accuracy is at 68.95%. In the TF category, the accuracy of each algorithm is as follows, neural network at 61.70%, decision tree at 61.68%, Naïve Bayes at 61.48%, rule induction at 61.12% and k-nearest neighbor at 28.68%. After combining it into an ensemble model, the accuracy is 61.40%. In the MF category, only the k-nearest neighbor has an accuracy of 83.21% while the rest and the ensemble model share an accuracy of 83.62%
3. Considering parameter setting: From the experiment, we found that the optimal K value for the k-nearest neighbor is 7 in the MM, TF, and MF categories, while the optimal k-value in the LR category is 5. Decision tree and rule induction algorithm attain the highest accuracy when setting criterion parameter to accuracy except for the TF category which setting the parameter to information gain yields a better result. The hidden layer of the neural network is set as 1 layer (9 nodes) for LR, 1 layer (9 nodes) for MM, 2 layers (9 and 8 nodes) for TF, and 4 layers (9, 8, 7, and 6 nodes) for MK when preprocessed with the first approach while setting the hidden layer after preprocessing by the second approach as follows: 4 Layers (6, 6, 5 and 5 nodes) for LR, 3 layers (6, 5 and 4 nodes) for MM, 2 layers (6 and 5 nodes) for TF and 1 layer (6 nodes) for MK
4. Considering fund category prediction: We developed this model to help customers pick suitable fund categories and risk levels. After we developed the ensemble model, we found that the neural network attains 93.43% accuracy while the ensemble model only attains 92.38%, in Fig. 8. However, to get the most accurate results, the data needs to be up-to-date, sufficient, and larger than the sample data that we used in this study. It would be more versatile if there were more fund categories. In the case of this study, neural networks can outperform the ensemble model, but if the dataset is changed, it is worth considering applying the ensemble model in that case in the future

Acknowledgment

The authors deeply acknowledge the insightful comments and suggestions made by the editors and reviewers of this manuscript. We express our thanks to Mahasarakham University for their support.

Funding Information

This research was financially supported by Mahasarakham University, Thailand.

Author's Contributions

Thongchai Kaewkiriya: System designed, data cleaning, literature survey, designed of the methodology, and written the manuscript.

Kittipol Wisaeng: Supervision, written, review, and edited.

Ethics

This article is an original research work. The corresponding author confirms that all of the other authors have read and approved the manuscript and no ethical issues are involved.

References

Ahmad, A., & Dey, L. (2007). A k-mean clustering algorithm for mixed numeric and categorical data. *Data & Knowledge Engineering*, *63*(2), 503-527. https://doi.org/10.1016/j.datak.2007.03.016

Anthony, M., & Bartlett, P. L. (2017). *Neural network learning: Theoretical foundations* (Vol. *9*). Cambridge: Cambridge university press. https://idea-stat.snu.ac.kr/book/2017%20neural%20network/20170814/ch8~11.pdf

Berwind, K., Bornschlegl, M., Hemmje, M., & Kaufmann, M. (2016). Towards a cross-industry standard process to support big data applications in virtual research environments. *CERC2016*, 82.

Cohen, W. W. (1995). Fast effective rule induction. In *Machine Learning Proceedings 1995* (pp. 115-123). Morgan Kaufmann. https://doi.org/10.1016/B978-1-55860-377-6.50023-2

Dietterich, T. G. (2000). Ensemble methods in machine learning. In *Multiple Classifier Systems: First International Workshop, MCS 2000 Cagliari, Italy, June 21–23, 2000 Proceedings 1* (pp. 1-15). Springer Berlin Heidelberg. https://doi.org/10.1007/3-540-45014-9_1

Faggiolani, C. (2011). Perceived identity: Applying grounded theory in libraries. *Perceived Identity: Applying Grounded Theory in Libraries*, 1-33. https://www.torrossa.com/it/resources/an/4383166

Freitas, A. A. (2002). Data mining and knowledge discovery with evolutionary algorithms. *Springer Science & Business Media*. ISBN-10: 9783540433316

Hastie, T., Tibshirani, R., Friedman, J. H., & Friedman, J. H. (2009). *The elements of statistical learning: Data mining, inference and prediction* (Vol. *2*, pp. 1-758). New York: Springer. https://doi.org/10.1007/978-0-387-21606-5

Hopfield, J. J. (1982). Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences*, *79*(8), 2554-2558. https://doi.org/10.1073/pnas.79.8.2554

Hulett, C., Hall, A., & Qu, G. (2012, August). Dynamic selection of k-nearest neighbors in instance-based learning. In *2012 IEEE 13th International Conference on Information Reuse & Integration (IRI),* (pp. 85-92). IEEE. https://doi.org/10.1109/IRI.2012.6302995

He, B., Shi, Y., Wan, Q., & Zhao, X. (2014). Prediction of customer attrition of commercial banks based on SVM model. *Procedia computer science*, *31*, 423-430. https://doi.org/10.1016/j.procs.2014.05.286

Jaiswal, D. P., Kumar, S., & Mukherjee, P. (2020). Customer transaction prediction system. *Procedia Computer Science*, *168*, 49-56. https://doi.org/10.1016/j.procs.2020.02.256

Kotsiantis, S. B., Zaharakis, I., & Pintelas, P. (2007). Supervised machine learning: A review of classification techniques. *Emerging Artificial Intelligence Applications in Computer Engineering*, *160*(1), 3-24. ISBN-10: 9781586037802

MacKay, D. J. (2003). *Information theory, inference and learning algorithms*. Cambridge university press.

Manish, K. (2012). Ensemble Techniques for Weak Learners. http://manish2020.blogspot.com/2012/12/ensemble-of-weak-learners.html

Murty, M. N., & Devi, V. S. (2011). *Pattern recognition: An algorithmic approach*. Springer Science & Business Media. ISBN: 9780857294951

Piryonesi, S. M., & El-Diraby, T. E. (2020). Role of data analytics in infrastructure asset management: Overcoming data size and quality problems. *Journal of Transportation Engineering, Part B: Pavements*, *146*(2), 04020022. https://doi.org/10.1061/JPEODX.0000175

Rolim, V., Mello, R. F., Nascimento, A., Lins, R. D., & Gašević, D. (2021, July). Reducing the size of training datasets in the classification of online discussions. In *2021 International Conference on Advanced Learning Technologies (ICALT)* (pp. 179-183). IEEE. https://doi.org/ 10.1109/ICALT52272.2021.00061

Sangsoi, C. (2015). The relationship between financial ratios and stock price of companies listed in the stock market of Thailand: A case study of service sector. *Master of Science (Finance) Independent Study, Bangkok University. [in Thai]*. https://so06.tci-thaijo.org/index.php/wms/article/view/245356

Sapaphan, A., (2016). Design decision support system for investment funds a case study of asset management (Thesis). *Thai-Nichi institute of technology*.

Sungkaew, J., (2001). Investment, 4th ed. *Thammasat University., Bangkok*.

Tanizaki, T., Hoshino, T., Shimmura, T., & Takenaka, T. (2020). Restaurants store management based on demand forecasting. *Procedia CIRP*, *88*, 580-583. https://doi.org/10.1016/j.procir.2020.05.101

Tao, T., Yan, K., & Yang, S. (2019, May). Classification of mutual fund investment types with advanced machine learning models. In *2019 IEEE International Conference on Big Data, Cloud Computing, Data Science & Engineering (BCD),* (pp. 84-89). IEEE. https://doi.org/10.1109/BCD.2019.8885073

Zhang, C., Zhang, H., & Wang, J. (2018). Personalized restaurant recommendation method combining group correlations and customer preferences. *Information Sciences*, *454*, 128-143. https://doi.org/10.1016/j.ins.2018.04.061