Predicting Startup Success Using Machine Learning Approach

Icha Wahyu Kusuma Ningrum [1]*, Farid Ridho [2], Arie Wahyu Wijayanto [3]****
* Department of Statistics, Politeknik Statistika STIS, Jakarta, Indonesia
** Department of Statistical Computing, Politeknik Statistika STIS, Jakarta, Indonesia
212011414@stis.ac.id [1], faridr@stis.ac.id [2], ariewahyu@stis.ac.id [2]

Article Info

Article history:

Received 2024-08-18
Revised 2024-08-27
Accepted 2024-09-13

Keyword:

Gradient Boosting,
K-Nearest Neighbor,
Random Forest,
Startup,
Support Vector Machines.

ABSTRACT

Predicting startup success is important because it helps investors, entrepreneurs, and stakeholders allocate resources more efficiently, minimize risks, and enhance decision-making in an uncertain and competitive environment. Therefore, investors need to predict whether a startup will succeed or fail. Investors conduct this assessment to determine if a startup is worthy of funding. The company's founders mark success here by receiving a sum of money through the Initial Public Offering (IPO) or Merger and Acquisition (M&A) process. If the startup closes, we will consider it a failure. The data used consists of 923 startup companies in the United States. We carried out the classification using four methods: Random Forest, Support Vector Machines (SVM), Gradient Boosting, and K-Nearest Neighbor (KNN). We then compare the results from the four methods with and without feature selection. We determine the feature selection based on the relative importance of each method. The results of this study indicate that the Random Forest method with feature selection has the best accuracy, precision, recall, and F1 score than the other methods, respectively 81.85%, 80.19%, 87.09%, and 83.44%.

I. INTRODUCTION

Startups make a significant impact on a nation's economy. Startup enterprises in the economy can foster entrepreneurial drive and encourage ongoing competition [1]. Conversely, startups have the capacity to generate novel ideas and address societal issues. Furthermore, this company's presence has the potential to create employment opportunities. Startups are a significant subject in economic policy for both developed and developing countries. They not only contribute to economic improvement but also have a profound effect on creativity and technical advancement [2].

Startups play a significant role in the economy, but they also come with a substantial amount of uncertainty and risk. According to [3], 90% of startups experience failure during their first year, and less than 40% of the remaining 10% manage to survive in the subsequent five years. Furthermore, according to [4], a significant 60% of businesses are unable to sustain themselves for a period of five years from their inception. Additionally, a staggering 75% of startups that secure investment ultimately meet with failure. Hence, accurately forecasting the viability of startups holds significant importance, particularly for investors. Investors anticipate a profit in exchange for the capital they invest in startups. When a startup fails, it will have a detrimental impact on investors. Using this forecast, investors can evaluate the viability of providing financial support to a firm.

Establishing a startup involves developing and introducing a novel product or service, which often entails inherent risks and uncertainties [5]. Startups often go through three fundamental stages in their life cycle: the starting stage, the expansion stage, and the maturity stage [6]. During the starting stage, a startup is a company with limited resources that endeavors to identify market problems, ascertain demand, and provide solutions to those problems. During the expansion stage, the organization experiences a period of rapid growth, with monthly growth rates reaching double digits. In the maturity stage, the startup has stabilized and undergone evaluation or measurement.

A starting business must consider uncertainty as a crucial factor. [5] assert that startups often face a significant level of uncertainty when developing new services or products. Startups in the technology industry encounter fierce competition and operate within a volatile and unpredictable environment [7]. Hence, effectively handling risk and uncertainty holds significant importance in the startup business, particularly for companies, investors, and venture capital.

Startup founders, family, friends, angel investors, and crowdfunding (fundraising) are common sources of initial funding, also known as seed funding. Moreover, one can also leverage formal funding sources such as banks, venture capital companies, and the government to supplement funding. Nevertheless, angel investors are the primary source of financing in the startup industry [8]. An angel investor is an individual who voluntarily allocates their funds to a startup in order to assist in its initial funding. After securing seed funding, the company continues to receive funding through rounds A, B, and beyond. Venture capital (VC) may provide funding to businesses during this phase. A venture capital company specializes in investing in entrepreneurs. The initial transfer of ownership of the company to external investors initiates this round A financing.

[5] used a machine learning approach to analyze data from 218,207 companies on Crunchbase from January 2011 to July 2021 in order to predict the success of startups. The science of machine learning explores methods for spotting patterns in large datasets and deriving knowledge from them. The four primary categories of machine learning are supervised learning, unsupervised learning, semi-supervised learning, and reinforcement learning [9]. The study [5] exclusively used the supervised learning algorithm, conducting the learning process based on the value of the objective variable, which the predictor variable influences. Consequently, the dataset utilized contains identifiers or classes. Logistic Regression, Decision Tree, Random Forest, Gradient Boosting, SVM, and Naïve Bayes comprise the methodologies implemented in the investigation. Additionally, the research did not attempt to classify based solely on the variable importance derived from each method.

Next, [10] conducted an additional study that sought to predict the success of startups using KNN, Decision Tree, and Naïve Bayes. The results of this study were 66.69%, 79.29%, and 64.21%, respectively. We utilized 19 out of the 49 features in the dataset. This investigation implemented data preprocessing to mitigate the presence of absent values. Furthermore, [11] employed the SVM method to forecast the success of a fledgling company by analyzing its acquisition status. The hyperplane value Kernel: Linear & C: 1.0 yielded an accuracy value of 79.1% in their investigation.

This study will employ a variety of variables. We select the employed variables from a variety of studies available in the Kaggle dataset to provide support. Firstly, we consider variables that are related to the funding that entrepreneurs utilize. A venture's funding status significantly influences its success [12]. Early-stage funding (seed funding) can establish a foundation for startup growth and further accelerate the business. Startups can also use early-stage funding to showcase their potential for success [13]. Secondly, previous research has shown that the type of investor can influence a venture's success. [5] observed that venture capital firms possess bargaining power over their portfolios. According to [14], the category of investor can significantly influence the valuation of a startup. Third, there is prior research that utilizes fundamental information about a company to forecast the success of a startup in its initial phases. This information includes the abilities and skills of the founders [15], the characteristics of the team and employees [14], and the location of the headquarters [15]. Fourth, we examined variables associated with the startup category. Researchers have identified industry characteristics as a significant factor in the success or failure of startups [16]. In certain sectors, startups may generate greater profitability than in others [5]. Therefore, a variety of industry characteristics can influence the success of startups.

Although numerous studies [5], [10], and [11] have attempted to compare various classification methods, none have yet conducted a classification by comparing the significance of variables in determining a startup's success in each machine learning method. Four critical reasons make feature selection essential. Firstly, choose to spare the model in order to minimize the number of parameters. Secondly, aim to reduce the training duration, alleviate overflow through enhanced generalization, and avoid the issue of dimensionality. Within the domain of data processing and analysis, the dataset can consist of a substantial number of factors or features that dictate the suitability and usefulness of the data [17]. Furthermore, the difficulty in classification lies in the need to include both balanced and imbalanced data [18]. Another incentive is to obtain the optimal model with accurate predictions and minimal mistakes [19]. Feature selection (FS) is the process of reducing the original feature set to a smaller subset while retaining the important information while rejecting the unnecessary ones [20].

Here is a summary of the study's key findings: Initially, it scrutinizes multiple characteristics to determine the ones that are valuable, specifically for the examination of classification data. Additionally, the system presents a comparison of various machine learning models, including Random Forest (RF), Support Vector Machine (SVM), K-Nearest Neighbor (KNN), and Gradient Boosting (GB), based on their crucial properties. Various models will have distinct capabilities in data categorization, which will impact the effectiveness of classification. In addition, we utilize the *varImp()* method for feature selection. In addition, our primary focus is on evaluating the feature selection application. We provide a description, analysis, and ideas for further study. Therefore, this investigation aims to improve the accuracy, precision, recall, and F1 score of startup success predictions. This is achieved by utilizing classification methods such as single models (SVM and KNN) and ensemble models (Random Forest and Gradient Boosting), while also considering variable importance within the context of startup data from the United States. Consequently, investors can mitigate the risk of losing their investment in a startup by utilizing the most effective classification method to foretell its success.

II. Methods

A. Startup's Success

The success of a startup is defined as an event that provides the company's founders with a substantial sum of money through the merger and acquisition (M&A) or initial public offering (IPO) procedure. If a company requires closure, it will be considered a failure. The success of a start-up is typically defined as a two-pronged approach. The company has the option of either conducting an IPO by listing its shares on a public stock market, thereby enabling its shareholders to sell them to the public, or acquiring or merging with another company, thereby providing those who have previously invested with immediate cash in exchange for their shares. People frequently refer to this procedure as an exit strategy [21].

Corporate restructuring frequently uses M&As. [22]define a merger as the process of merging two companies to create a single entity, typically under a new name, with the objective of enhancing the company's profitability and sales. This strategy is more common among non-tech companies of comparable size and status. M&A activities are particularly crucial for high-tech industries, as they frequently employ M&As to acquire cutting-edge technologies or rapidly expand their R&D capabilities [23]. An acquisition is a scenario in which one organization acquires another, resulting in the latter's demise. An acquisition is the process by which one organization acquires a dominant interest in another organization.

In order to forecast the performance of a startup, it is essential to establish a definition of success. [24] observed that the public status of a startup is a critical criterion for success and substantially influences the investment decisions of a VC firm. [25] used IPOs as a valuable indicator of startup success and demonstrated that they are the most significant factor in determining VC investments in startups. According to [26], an IPO is the most significant criterion for success because of its transparency and high availability of information.

As previously stated, an effective initial public offering (IPO) indicates that investors in the stock exchange market find a company intriguing [15]. A successful initial public offering (IPO) is indicative of a company's market presence and likelihood of survival. Therefore, we identify an IPO as an output variable that signifies a company's prosperity.

B. Random Forest

Decision trees employ the Random Forest approach to classify data. This approach involves k randomly generated trees that are mutually exclusive [27]. The benefits of Random Forest are evident in its strong performance and straightforward structure. Therefore, we utilize this technique for both categorization and prediction. The Random Forest algorithm is shown below.

1. Creating new training sets by randomly selecting samples from the existing training set, allowing for duplicates (bootstrap).
2. We construct a tree for each fresh training set, using random feature selection at each node and without any pruning. We employ the same approach for each individual tree when constructing the CART (Classification and Regression Tree).
3. Once we have constructed a substantial number of trees, we will forecast fresh data by aggregating the outcomes of each tree through majority voting.

Figure 1 illustrates the Random Forest algorithm.

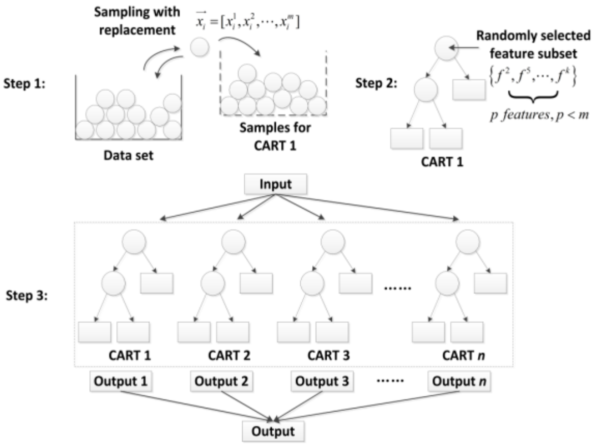
Sampling with replacement
$\vec{x}_i = [x_i^1, x_i^2, \cdots, x_i^m]$
Step 1:
Data set
Samples for CART 1
Step 2:
Randomly selected feature subset
$\underbrace{\{f^2, f^5, \cdots, f^k\}}_{p\ features,\ p<m}$
CART 1
Input
Step 3:
CART 1
CART 2
CART 3
......
CART *n*
Output 1
Output 2
Output 3
......
Output *n*
Output

Figure 1. Random Forest Algorithm Flowchart [28]

C. Support Vector Machines (SVM)

Support Vector Machines (SVM) is a highly effective technique for classification and regression tasks [29]. We employ a SVM to identify an ideal hyperplane in the crystal structure that effectively separates different classes. The SVM parameters govern crystal function normalization. [30] assert an inverse relationship between the value and the normalization intensity. The subfactors enhance the classifier's crystal function and the margin of the crystal surface. This can lead to the creation of models that accurately match the data. Models with high values of C parameters have a tendency to correctly categorize all training samples.

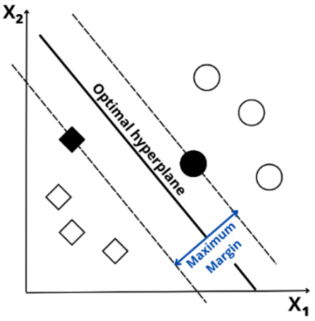

Figure 2. Maximum Margin in Hyperplane Determination

The kernel concept allows SVM to be applied to nonlinear data. There exist three kernel functions:

Kernel Polynomial with degree h

$$K(X_i, X_j) = (X_i . X_j + 1)^h \quad (1)$$

Kernel Gaussian Radial Basis Function

$$K(X_i, X_j) = e^{|X_i - X_j|^{2/2\sigma^2}} \quad (2)$$

Kernel Sigmoid

$$K(X_i, X_j) = \tanh(kX_i . X_j - \delta) \quad (3)$$

D. Gradient Boosting

Gradient Boosting is a machine learning technique that constructs a collection of decision trees. Every subsequent tree is constructed taking into account the vulnerabilities of the preceding tree. Gradient Boosting fundamentally posits that merging the subsequent model with the previous one will minimize the overall prediction error [31].

E. K-Nearest Keigbor (KNN)

Commonly used for distance-based classification, the K-Nearest Neighbor (KNN) technique is renowned for its ease of implementation and extensive applicability. In this context, the Euclidean distance metric is used. The typical steps involved in the K-Nearest Neighbors (KNN) algorithm are as follows:

1. Determine the value of the parameter (K) for class determination.
2. Calculate the distance between the new data and each point in the dataset.
3. Select a set of K data points with the shortest distance, and then classify the new data point.

F. Imbalance Data

The model may exhibit a bias towards classifying instances into the majority class due to imbalanced data, which requires attention. We can employ various methods to address unbalanced data, including adaptive synthetic (ADASYN), synthetic minority oversampling technique (SMOTE), and majority-weighted minority oversampling technique (MWMOTE). However, this study will utilize MWMOTE to manage data in the event of imbalance. According to [32], MWMOTE is an enhancement of the SMOTE method that involves weighting and grouping the synthetic data generated for minority data. Moreover, we shall acquire representative synthetic data. The outcomes of this procedure can mitigate bias and overfitting, resulting in synthetic data that exhibits a higher level of accuracy through the clustering process.

G. K-Fold Cross Validation

The K-fold cross validation method evaluates a model's performance. This approach involves partitioning the dataset into two subsets: one for training and the other for testing. In this approach, we select the training data in a more organized and systematic manner, leading to a clear contrast. Presented below is the algorithm for cross-validation.

1. Partition the data into k subgroups of equal magnitude.
2. Use each subset as testing data, with the remaining subsets as training data up to the kth fold.
3. Next, compute the sum of the k components' accuracies and divide the result by k. Consequently, we calculate the mean accuracy, which then becomes the ultimate precision.

H. Feature Selection

Feature selection is the process of removing excessive or irrelevant features that are not pertinent to the prediction task at hand. Implementing feature selection techniques can effectively decrease the computational time required and enhance evaluation metrics, such as accuracy. The inclusion of irrelevant features in the research will reduce accuracy [33]. [34] used the *varimp()* function to identify the significant variables in the model. The study incorporates the wrapper technique to pick significant variables, as it does feature selection concurrently with the modeling implementation. Feature selection is beneficial across various areas, including ecology, climate, health, and finance. The evaluation of a function's variable and feature importance is contingent upon whether the model utilizes information or not. The main benefit of employing a model-based approach is its strong association with the model's performance and its ability to integrate the correlation structure among predictors into the relevance calculation. Clearly, the significance is computed. Each predictor will have a unique significance variable for each class. We then rescale all the significant measurements to a maximum value of 100.

Varimp()'s practical application closely resembles the Random Forest method. The experiment utilized the Random Forest model from the R package to measure model-specific metrics. We record the prediction accuracy for each tree on a specific section of the data. After permuting each predictor variable, we complete the process. We then calculate the difference between the two accuracys as an average across all trees, and adjust it by normalizing it with the standard error. Next, employ the *varimp()* function to ascertain the significance of each aspect.

Prior research demonstrating varimp's efficacy and efficiency in identifying significant features necessitates the exclusion of all features in the dataset. As a result, it will impact the duration of processing, perhaps leading to improved accuracy and more features associated with higher-dimensional data.

I. Model Evaluation

The metrics utilized in this study encompass accuracy, precision, recall, and F1 score, as represented by the following equation. We derive the score from the confusion matrix. The confusion matrix is a tabular representation that effectively describes the performance of a classification model.

TABLE I. CONFUSION MATRIX

		Prediction	
		Positive	**Negative**
Actual	**Positive**	True Positive (TP)	False Negative (FN)
	Negative	False Positive (FP)	True Negative (TN)

$$Akurasi = \frac{TP+TN}{TP+FP+TN+FN} \quad (4)$$

$$Precision = \frac{TP}{TP+FP} \quad (5)$$

$$Recall = \frac{TP}{TP+FN} \quad (6)$$

$$F1\ score = \frac{2 \times Precision \times Recall}{Precision+Recall} \quad (7)$$

Accuracy is a metric that quantifies the degree to which a model makes valid predictions. We determine accuracy by dividing the count of accurately classified observations by the total count of data. Precision gauges the precision of accurately forecasting positive data from all projected positive data. Recall, or sensitivity, quantifies the accuracy of properly predicting positive data among the actual positive data. The F1 Score is used to measure the weighted average of precision and recall. When there is a significant disparity or distance between the number of false positives and false negatives in the analyzed data, the F1 score performs most effectively. If the false positive and false negative values are about equal (symmetric), then accuracy is a more suitable metric for evaluating the model.

J. Data

We obtained the data for this investigation from the Kaggle website. The startup success prediction dataset comprises 923 observations of enterprises in the United States from 2005 to 2012. This dataset comprises 49 features, with 48 features serving as attributes and one feature serving as a class or label (acquired/closed).

In this study, the selection of variables is based on a review of previous literature and adjusted to the available dataset. There are several categorical variables used in this study, namely Is_CA (Is the startup company in California?), Is_NY (Is the startup company in New York?), Is_MA (Is the startup company in Massachusetts?), Is_TX (Is the startup company in Texas?), is_otherstate, Category_code (Category of the field that the startup is engaged in), is_software, Is_web, Is_mobile, Is_enterprise, Is_advertising, Is_gamesvideo, Is_commerce, Is_biotech, Is_consulting, Is_othercategory, Has_VC (has venture capital), Has_angel (has angel investors), Has_roundA, Has_roundB, Has_roundC, Has_roundD, and Is_top500. In addition, there are also several numeric variables used in the study including Latitude, Longitude, Age_first_funding_year, Age_last_funding_year, Relationships, Funding_rounds, Funding_total_usd, Milestones, and Avg_participants. We will use these variables to predict startup success (acquired/closed).

K. Analysis Steps

The first step in the analysis involves data preparation and variable selection, which is based on the findings of the literature review. Next, we conduct data preprocessing, which includes tasks like data transformation and rectifying anomalous results. Apply the appropriate classification modeling approach before proceeding to the modeling stage. This study involved two rounds of modeling. The first stage utilized the variables specified at the beginning, based on the literature review. The second stage utilized only the significant factors determined by each respective approach. Moreover, we conducted a comprehensive assessment to determine the efficacy of the employed methodology.

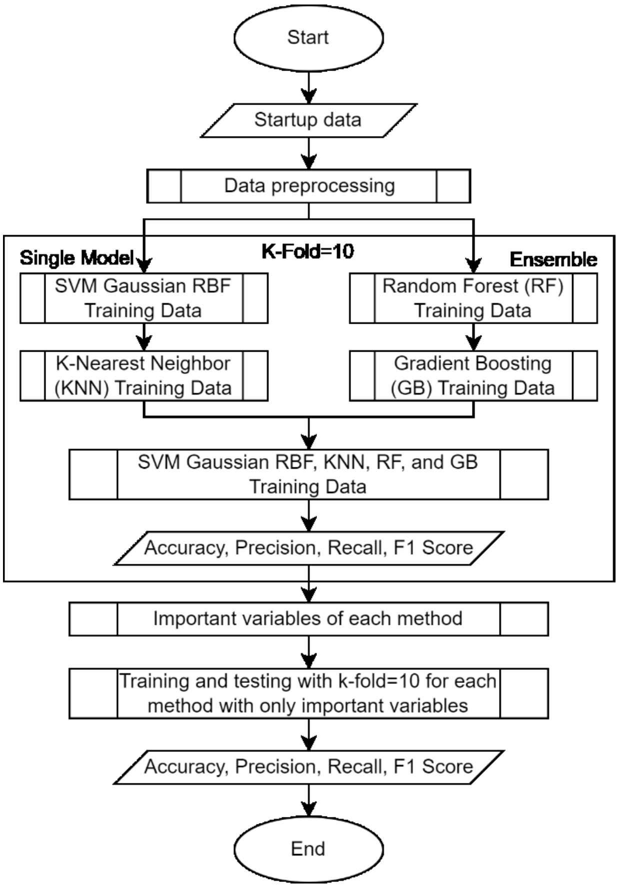
Start
Startup data
Data preprocessing
Single Model
K-Fold=10
Ensemble
SVM Gaussian RBF Training Data
Random Forest (RF) Training Data
K-Nearest Neighbor (KNN) Training Data
Gradient Boosting (GB) Training Data
SVM Gaussian RBF, KNN, RF, and GB Training Data
Accuracy, Precision, Recall, F1 Score
Important variables of each method
Training and testing with k-fold=10 for each method with only important variables
Accuracy, Precision, Recall, F1 Score
End

Figure 3. Research Flow Chart

This study employed machine learning models such as Support Vector Machines (SVM), K-Nearest Neighbors (KNN), Random Forests, and Gradient Boosting. Prior research by [35], which forecasts the success of startups, informs the choice of these strategies based on their respective benefits. Furthermore, [36] conducted a study that employed various machine learning approaches, such as K-Nearest Neighbor (KNN) and SVM, to predict the performance of startups. The research involved surveying 265 information and communication technology (ICT) enterprises in Australia. Furthermore, they employed the machine learning decision tree technique alongside these methodologies. The study excluded decision trees due to the superior classification performance of the Random Forest method. Random Forest (RF) is an ensemble classifier model that learns from decision tree models (DT) and outperforms them. Errors in the middle of the process do not affect RF, unlike DT, as it does not propagate them to subsequent steps.

III. RESULT AND DISCUSSION

A. Data Preprocessing

The initial phase involves examining the current data structure. Modify the inappropriate data type to make it suitable. Additionally, we conduct a descriptive analysis to spot patterns in the data and pinpoint any anomalous entries. Figure 4 shows that investors initially funded the majority of businesses between the ages of 0 and 2. Furthermore, we discovered an inconceivably negative age. We will address this issue later during the data preparation phase. Examining the distribution of acquired and closed startups reveals virtually no difference in their initial funding age, irrespective of acquisition or closure.

Figure 5 shows that the venture was primarily 3-4 years old when investors last funded it. Furthermore, we discovered an inconceivably negative age. We will address it later during the data preparation phase. The distribution of acquired and closed enterprises reveals that the latter were smaller in age at the time of their most recent funding than the former. This implies that the startups that failed (closed) were relatively young at the time of their last funding, or it could also indicate that no investors were considering funding them for an extended period.

Figure 6 shows a skewed distribution of relationships to the right. This suggests that startups are exceedingly numerous, while businesses with many relationships are exceedingly scarce. Both acquired and closed startups exhibit an identical distribution of relationships, both skewed to the right. However, some startups with the closed label have fewer relationships compared to those with the acquired label. The total funding (USD) is significantly right-skewed, as evidenced by Figure 7. The startup's total funding ranges from a minimum of 11,000 USD to a maximum of 57,000,000,000 USD, resulting in an average of 2,542,000 USD.

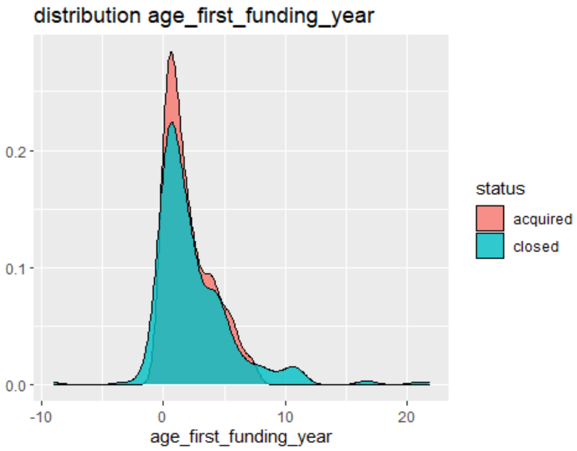

Figure 4. Distribution of the variable age_first_funding _year according to startup status (acquired/closed)

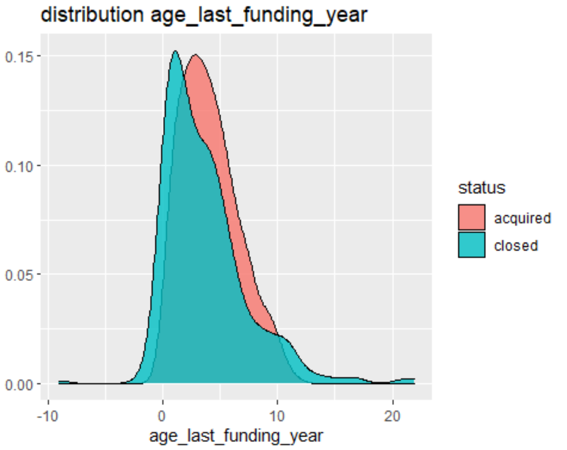

Figure 5. Distribution of the variable age_first_funding _year according to startup status (acquired/closed)

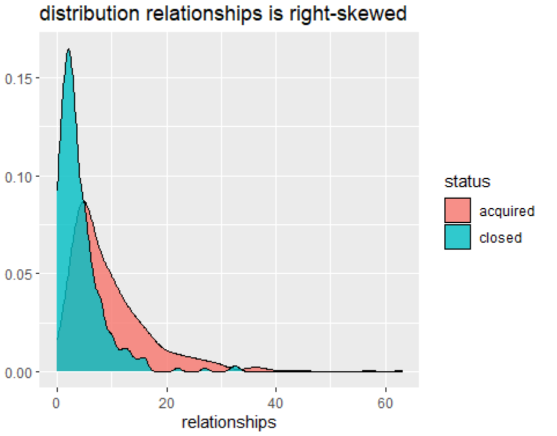

Figure 6. Distribution of relationships variables by startup status (acquired/closed)

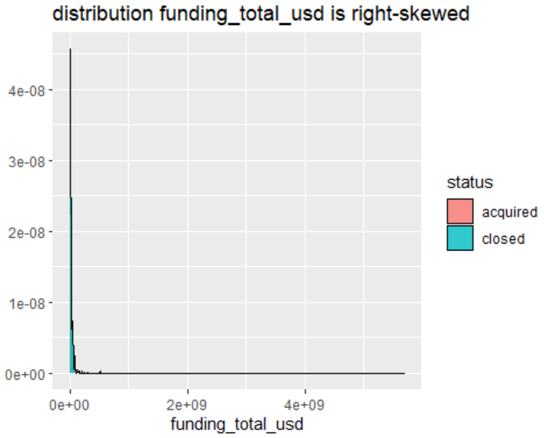

Figure 7. Distribution of funding_total_usd variable by startup status (acquired/closed)

The dataset used did not contain any missing entries. The descriptive analysis revealed anomalous entries, specifically negative values in the variables age_first_funding_year and age_last_funding_year. Consequently, the researcher will rectify this unusual result using the K-Nearest Neighbor (KNN) algorithm by replacing missing values with negative values and then imputing.

In 46 instances, the variable age_first_funding_year has a negative value. There are 13 records with a negative value for the age_last_funding_year variable. After the imputation process, the variable no longer contains any negative entries. Examining the dataset by state (acquired/closed) revealed imbalance. Consequently, it requires proper management.

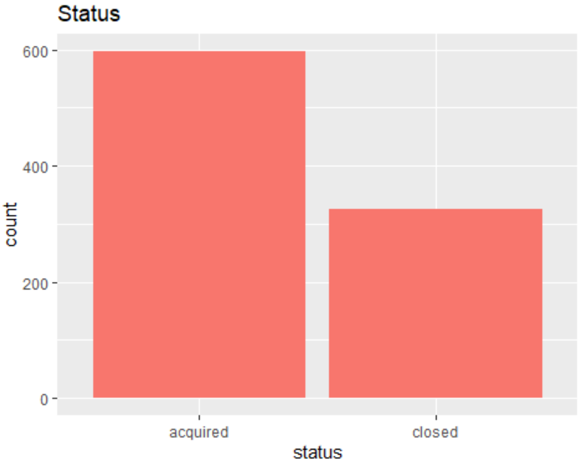

Figure 8. Plot number of observations by status

Figure 8 clearly shows that it is not proportionate. The imbalance ratio is 0.546. There are 597 companies with acquired status and 326 companies with closed status among the 923 companies in the data. The researcher employs the MWMOTE technique to manage imbalance data, achieving a ratio of 0.90. Figure 8 illustrates the number of observations after processing.

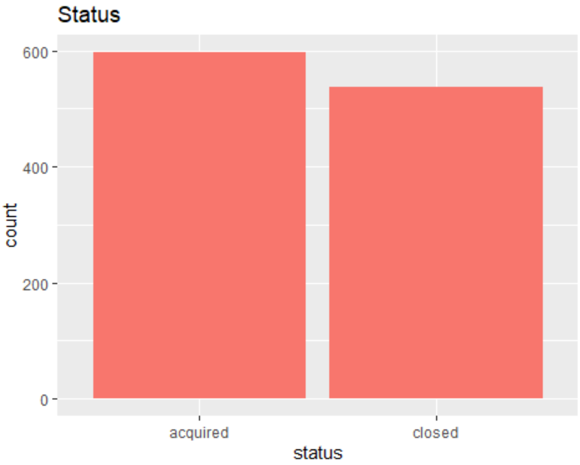

Figure 9. Plot of the number of observations by status after handling data imbalance using MWMOTE

B. *Model Evaluation*

. Table II shows the comparative evaluation of the models.

TABLE II. COMPARISON OF MODEL EVALUATION

Model	**Accuracy**	**Precision**	**Recall**	**F1 Score**
Random Forest	0.811	0.799	0.859	0.827
SVM	0.789	0.775	0.846	0.808
Gradient Boosting	0.801	0.799	0.832	0.814
KNN	0.599	0.615	0.643	0.627

In terms of accuracy, precision, recall, and F1 score values, Table II shows that the Random Forest algorithm model does better than the SVM, Gradient Boosting, and KNN models. [37] confirms the effectiveness of the Random Forest approach. Table II reveals that the accuracy and precision values between the Random Forest and Gradient Boosting approaches exhibit minimal disparity. Afterwards, we found that the KNN approach outperforms other methods in terms of accuracy, precision, recall, and F1 score.

C. *Variable Importance*

Each method employed in the categorization process yields significant variables. The relevance variable quantifies the importance of a predictor variable in the categorization process. The significance score reveals this. A higher score indicates a greater significance of the predictor variable. Figures 10, 11, 12, and 13 display the variable important scores for classifying startup success using the Random Forest, SVM, Gradient Boosting, and KNN approaches. We derive the importance scores sequentially, starting from the highest and moving towards the lowest.

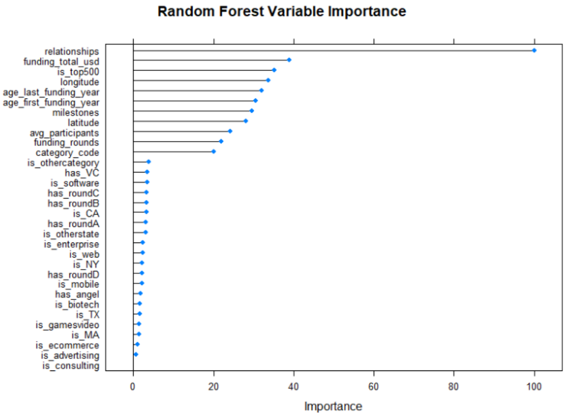

Figure 10. Plot the importance of the variable by its score using the Random Forest Method.

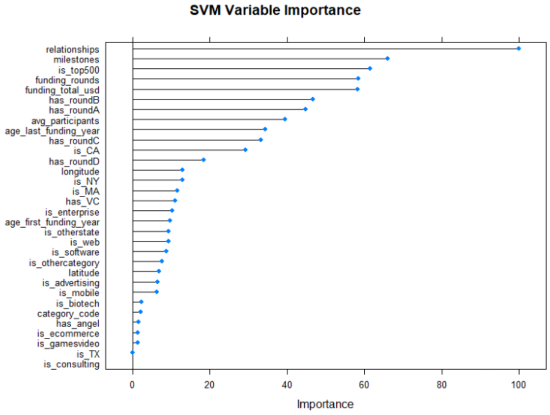

Figure 11. Plot the importance of the variable by its score using the SVM Method.

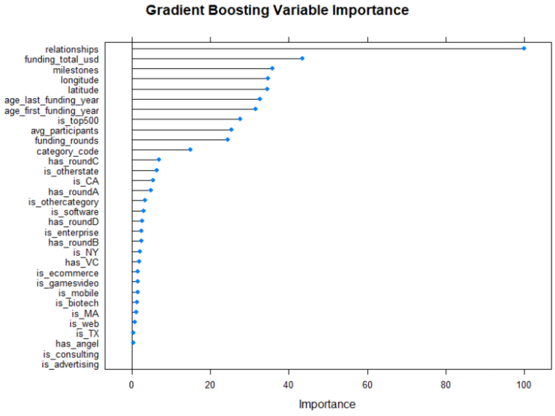

Figure 12. Plot the importance of the variable by its score using the Gradient Boosting Method.

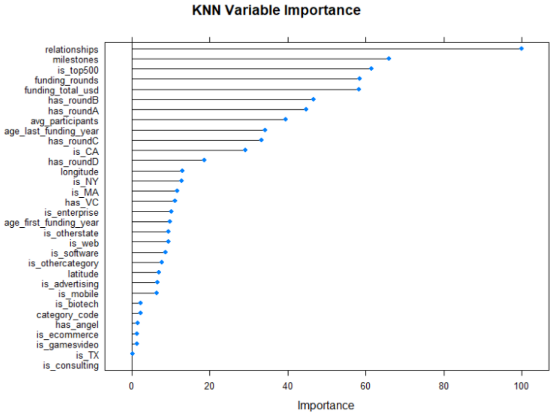

Figure 13. Plot the importance of the variable by its score using the KNN Method.

TABLE III. VARIABLES WITH HIGH IMPORTANTT SCORES IN THE RANDOM FOREST METHOD

Variables	Important Scores
relationships	100.000
funding_total_usd	39.000
is_top500	35.248
longitude	33.725
age_last_funding_year	32.089
age_first_funding_year	30.582
milestones	29.567
latitude	28.108
avg_participants	24.235
funding_rounds	21.869
category_code	20.126

Table III displays the variable importance in the Random Forest approach, arranged in descending order based on their importance scores. The variable is_othercategory has a significantly different relevance score compared to the category_code, which is 3.859. Therefore, we do not consider this variable to be important.

TABLE IV. VARIABEL DENGAN SKOR IMPORTANT TINGGI PADA METODE SVM RBF

Variables	Important Scores
relationships	100.000
milestones	66.038
is_top500	61.511
funding_rounds	58.436
funding_total_usd	58.270
has_roundB	46.644
has_roundA	44.833
avg_participants	39.484
age_last_funding_year	34.279
has_roundC	33.206
is_CA	29.193

Table IV displays the variable importance in the SVM RBF approach, ordered in descending order based on their importance scores. The variable has_roundD has a significantly different importance score compared to the variable is_CA, which is only 18.537. Therefore, we do not consider the variable has_roundD to be important.

TABLE V. VARIABLES WITH HIGH IMPORTANTT SCORES IN THE GRADIENT BOOSTING METHOD

Variables	Important Scores
relationships	100.000
funding_total_usd	43.556
milestones	35.748
longitude	34.660
latitude	34.446
age_last_funding_year	32.696
age_first_funding_year	31.575
is_top500	27.523
avg_participants	25.356
funding_rounds	24.336
category_code	14.998

Table V displays the variable importance in the Gradient Boosting approach, ranked by the highest important score. The variable has_roundC has a significantly different relevance score compared to the variable category_code, which is 6.868. Therefore, we do not consider the variable has_roundC to be important.

TABLE VI. VARIABLES WITH HIGH IMPORTANTT SCORES IN THE KNN METHOD

Variables	Important Scores
Relationships	100.000
Milestones	66.038
is_top500	61.511
funding_rounds	58.436
funding_total_usd	58.270
has_roundB	46.644
has_roundA	44.833
avg_participants	39.484
age_last_funding_year	34.279
has_roundC	33.206
is_CA	29.193

Table VI displays the KNN approach's variable importance, arranged in descending order based on their importance scores. The variable has_roundD has a relatively low relevance score of 18.537 when compared to the variable is_CA. Employing SVM with a radial basis function (RBF) kernel yields an identical outcome.

From this, it is evident that the classification of startup success is influenced by relationship variables and funding-related variables. This implies that the model predicting the success of a fledgling company heavily relies on these variables. One of the critical factors that determines the success of a startup is the number of primary business relationships or partnerships that a startup has. This is due to the fact that consumers are more likely to regard a startup as trustworthy and necessary as it establishes more business relationships. Prior research [38] has demonstrated that a startup's popularity, attractiveness, and exposure to consumers influence its success.

[3] have identified the status of funding as a critical factor in the success of startups. In particular, early-stage funding can serve as the foundation for the expansion of a startup, thereby accelerating its growth. Additionally, startups may utilize early-stage funding to exhibit their potential for success [13]. Funding status can serve as an effective predictor of success, as the majority of startups receive funding from venture capital firms and other sources [12].

D. Evaluation of the Variable Importance Model

We present the subsequent outcomes of the model evaluation based solely on the variable importance, following the used methodology.

TABLE VII. COMPARISON OF MODEL EVALUATION AFTER SELECTION FEATURES USING VARIABLE IMPORTANCE

Model	Accuracy	Precision	Recall	F1 Score
Random Forest	0.818	0.802	0.871	0.834
SVM	0.750	0.731	0.831	0.777
Gradient Boosting	0.796	0.798	0.822	0.809
KNN	0.606	0.617	0.674	0.642

The accuracy, precision, recall, and F1 score values produced by the Random Forest and KNN approaches rose after picking features based solely on variable importance. In contrast, the SVM and Gradient Boosting approaches showed a drop in performance when the classification solely relied on essential variables. [39] conducted research that elucidates this phenomenon. They found that reducing the number of features and using the Support Vector Machine (SVM) method can lead to a decrease in accuracy. This is because, although the features may have a minimal impact on the classification model, using a non-probabilistic algorithm like SVM significantly affects the overall accuracy.

Based on the performance evaluation results in our studies, it is evident that Random Forest is the superior classifier. The reason for this is that Random Forest utilizes the prediction outcomes of several decision trees through majority voting, resulting in more precise predictions. The K-Nearest Neighbors (KNN) method offers the advantage of low temporal complexity, enabling it to categorize data quickly in comparison to other machine learning methods. Nevertheless, it fails to take into account the minority class and the weight of data points, potentially leading to a decrease in accuracy for datasets with a high level of noise [40].

Moreover, the SVM model exhibits benefits when the data is inherently non-linear. We can employ SVM for non-linear classification scenarios by utilizing kernels such as the Radial Basis Function (RBF). Support Vector Machines (SVM) can be a valuable tool for predicting the success of startups when there is data irregularity, such as when the data is not evenly distributed or its distribution is uncertain [41]. Nevertheless, Support Vector Machine (SVM) exhibits reduced efficacy when confronted with noisy data because of its susceptibility to outliers and noise within the dataset. These factors may disrupt the hyperplane's positioning, resulting in decreased model accuracy.

Additionally, the versatility of Gradient Boosting allows for its application with various loss functions, providing great flexibility in addressing diverse classification challenges. Gradient Boosting is prone to overfitting if the parameters, particularly the number of trees and learning rate, are not appropriately configured. Gradient Boosting is prone to overfitting, particularly when used to small or noisy datasets [42].

IV. Conclusion

We can infer from the conducted analysis and discussions that the Random Forest approach outperforms the SVM, Gradient Boosting, and KNN methods in terms of accuracy, precision, recall, and F1 score. However, not all strategies used after feature selection result in improved model evaluations. These methods only consider variables that are believed to be relevant and have a high variable relevance score.

According to the research, investors who want to predict the success or failure of startups should take into account key elements that are believed to have an impact on their success. These variables include relationships and factors associated with startup funding. The appropriate classification method for forecasting startup success is Random Forest. Both comparison and reduction methods can perform feature selection, especially when dealing with a large number of features. Recommendations for future research include augmenting the dataset by incorporating more data sources to enhance the training data.

In the future, we intend to develop a distinct data set or data repository to ascertain the comparability of the results when applied to various datasets. This study, as previously indicated, employs data from startup companies in the United States. Consequently, there may be modifications necessary when the data is applied to other countries, such as the location variable (categorical) of startup companies. The location of a venture will influence its success.

Acknowledgment

We are appreciative of the generous support that Politeknik Statistika STIS has provided in funding this research.

References

[1] B. Kim, H. Kim, and Y. Jeon, “Critical Success Factors of a Design Startup Business,” *Sustainability*, vol. 10, no. 9, p. 2981, Aug. 2018, doi: 10.3390/su10092981.

[2] M. I. Luger and J. Koo, “Defining and Tracking Business Start-Ups,” *Small Business Economics*, vol. 24, no. 1, pp. 17–28, Jan. 2005, doi: 10.1007/s11187-005-8598-1.

[3] C. Unal and I. Ceasu, “A Machine Learning Approach Towards Startup Success Prediction,” 2019, [Online]. Available: http://irtg1792.hu-berlin.de

[4] C. Giardino, S. S. Bajwa, X. Wang, and P. Abrahamsson, *Agile Processes in Software Engineering and Extreme Programming*, vol. 212. in Lecture Notes in Business Information Processing, vol. 212. Cham: Springer International Publishing, 2015. doi: 10.1007/978-3-319-18612-2.

[5] J. Kim, H. Kim, and Y. Geum, “How to succeed in the market? Predicting startup success using a machine learning approach,” *Technol Forecast Soc Change*, vol. 193, Aug. 2023, doi: 10.1016/j.techfore.2023.122614.

[6] A. Skala, *Digital Startups in Transition Economies*. Cham: Springer International Publishing, 2019. doi: 10.1007/978-3-030-01500-8.

[7] S. Tomy and E. Pardede, “From Uncertainties to Successful Start Ups: A Data Analytic Approach to Predict Success in Technological Entrepreneurship,” *Sustainability*, vol. 10, no. 3, p. 602, Feb. 2018, doi: 10.3390/su10030602.

[8] M. S. Dewi and Kartini, “Angel Investor Investment Decision Making Criteria in Startup Business,” 2022. [Online]. Available: https://journal.uii.ac.id/selma/index

[9] G. Shobha and S. Rangaswamy, “Machine Learning,” 2018, pp. 197–228. doi: 10.1016/bs.host.2018.07.004.

[10] A. Prayoga Permana, K. Ainiyah, and K. Fahmi Hayati Holle, “Analisis Perbandingan Algoritma Decision Tree, kNN, dan Naive Bayes untuk Prediksi Kesuksesan Start-up,” 2021. [Online]. Available: https://www.kaggle.com/manishkc06/startup-success-prediction.

[11] A. E. Goldenia, C. Chairunnisa, H. Harisa, J. Christian, D. Desta, and S. Prasvita, *Implementation of Support Vector Machine Algorithm in Predicting Startup Success Based on Acquisition Status*. 2021.

[12] D. Camelo Martinez, “Startup Success Prediction in The Dutch Startup Ecosystem,” 2019. [Online]. Available: http://repository.tudelft.nl/.

[13] M. Islam, A. Fremeth, and A. Marcus, “Signaling by early stage startups: US government research grants and venture capital funding,” *J Bus Ventur*, vol. 33, no. 1, pp. 35–51, Jan. 2018, doi: 10.1016/j.jbusvent.2017.10.001.

[14] A. Köhn, “The determinants of startup valuation in the venture capital context: a systematic review and avenues for future research,” *Management Review Quarterly*, vol. 68, no. 1, pp. 3–36, Feb. 2018, doi: 10.1007/s11301-017-0131-5.

[15] K. Żbikowski and P. Antosiuk, “A machine learning, bias-free approach for predicting business success using Crunchbase data,” *Inf Process Manag*, vol. 58, no. 4, p. 102555, Jul. 2021, doi: 10.1016/j.ipm.2021.102555.

[16] C. Bandera and E. Thomas, “The Role of Innovation Ecosystems and Social Capital in Startup Survival,” *IEEE Trans Eng Manag*, vol. 66, no. 4, pp. 542–551, Nov. 2019, doi: 10.1109/TEM.2018.2859162.

[17] J. K. Jaiswal and R. Samikannu, “Application of Random Forest Algorithm on Feature Subset Selection and Classification and Regression,” in *2017 World Congress on Computing and Communication Technologies (WCCCT)*, IEEE, Feb. 2017, pp. 65–68. doi: 10.1109/WCCCT.2016.25.

[18] R.-C. Chen, “Using Deep Learning to Predict User Rating on Imbalance Classification Data,” *IAENG Int J Comput Sci*, 2019.

[19] L. A. García-Escudero, A. Gordaliza, C. Matrán, and A. Mayo-Iscar, “A review of robust clustering methods,” *Adv Data Anal*

Classif, vol. 4, no. 2–3, pp. 89–109, Sep. 2010, doi: 10.1007/s11634-010-0064-5.

[20] A. L. Blum and P. Langley, "Selection of relevant features and examples in machine learning," *Artif Intell*, vol. 97, no. 1–2, pp. 245–271, Dec. 1997, doi: 10.1016/S0004-3702(97)00063-5.

[21] B. Guo, Y. Lou, and D. Pérez-Castrillo, "Investment, Duration, and Exit Strategies for Corporate and Independent Venture Capital-Backed Start-Ups," *J Econ Manag Strategy*, vol. 24, no. 2, pp. 415–455, Jun. 2015, doi: 10.1111/jems.12097.

[22] A. Alam and S. Khan, "Strategic Management: Managing Mergers & Acquisitions," 2014.

[23] C.-P. Wei, Y.-S. Jiang, and C.-S. Yang, "Patent Analysis for Supporting Merger and Acquisition (M&amp;A) Prediction: A Data Mining Approach," 2009, pp. 187–200. doi: 10.1007/978-3-642-01256-3_16.

[24] S. J. Chang, "Venture capital financing, strategic alliances, and the initial public offerings of Internet startups," *J Bus Ventur*, vol. 19, no. 5, pp. 721–741, Sep. 2004, doi: 10.1016/j.jbusvent.2003.03.002.

[25] L. A. Jeng and P. C. Wells, "The determinants of venture capital funding: evidence across countries," *Journal of Corporate Finance*, vol. 6, no. 3, pp. 241–289, Sep. 2000, doi: 10.1016/S0929-1199(00)00003-1.

[26] T. E. Stuart, H. Hoang, and R. C. Hybels, "Interorganizational Endorsements and the Performance of Entrepreneurial Ventures," *Adm Sci Q*, vol. 44, no. 2, pp. 315–349, Jun. 1999, doi: 10.2307/2666998.

[27] L. Breiman, "Random Forests," 2001.

[28] L. Lin, F. Wang, X. Xie, and S. Zhong, "Random forests-based extreme learning machine ensemble for multi-regime time series prediction," *Expert Syst Appl*, vol. 83, pp. 164–176, Oct. 2017, doi: 10.1016/j.eswa.2017.04.013.

[29] C. J. C. Burges, "A Tutorial on Support Vector Machines for Pattern Recognition," *Data Min Knowl Discov*, vol. 2, no. 2, pp. 121–167, 1998, doi: 10.1023/A:1009715923555.

[30] F. Pedregosa *et al.*, "Scikit-learn: Machine Learning in Python," *J. Mach. Learn. Res.*, vol. 12, no. null, pp. 2825–2830, Nov. 2011.

[31] J. H. Friedman, "Greedy function approximation: A gradient boosting machine.," *The Annals of Statistics*, vol. 29, no. 5, Oct. 2001, doi: 10.1214/aos/1013203451.

[32] S. Barua, Md. M. Islam, X. Yao, and K. Murase, "MWMOTE--Majority Weighted Minority Oversampling Technique for Imbalanced Data Set Learning," *IEEE Trans Knowl Data Eng*, vol. 26, no. 2, pp. 405–425, Feb. 2014, doi: 10.1109/TKDE.2012.232.

[33] S. Doraisamy, S. Golzari, N. Norowi, md nasir Sulaiman, and N. Udzir, *A Study on Feature Selection and Classification Techniques for Automatic Genre Classification of Traditional Malay Music.* 2008.

[34] R.-C. Chen, C. Dewi, S.-W. Huang, and R. E. Caraka, "Selecting critical features for data classification based on machine learning methods," *J Big Data*, vol. 7, no. 1, p. 52, Dec. 2020, doi: 10.1186/s40537-020-00327-4.

[35] J. Kim, H. Kim, and Y. Geum, "How to succeed in the market? Predicting startup success using a machine learning approach," *Technol Forecast Soc Change*, vol. 193, p. 122614, Aug. 2023, doi: 10.1016/j.techfore.2023.122614.

[36] S. Tomy and E. Pardede, "From Uncertainties to Successful Start Ups: A Data Analytic Approach to Predict Success in Technological Entrepreneurship," *Sustainability*, vol. 10, no. 3, p. 602, Feb. 2018, doi: 10.3390/su10030602.

[37] Y. Aryani and A. W. Wijayanto, "Classification of Radar Returns from the Ionosphere Using SVM, Naive Bayes and Random Forest," *Komputika : Jurnal Sistem Komputer*, vol. 10, no. 2, pp. 111–117, Sep. 2021, doi: 10.34010/komputika.v10i2.4347.

[38] B. Chitkara and S. M. J. Mahmood, "Importance of Web Analytics for the Success of a Startup Business," 2020, pp. 366–380. doi: 10.1007/978-981-15-5830-6_31.

[39] A. Rahmansyah, O. Dewi, P. Andini, T. Hastuti, P. Ningrum, and M. E. Suryana, "Comparing the Effect of Feature Selection on the Naïve Bayes and Support Vector Machine Algorithms," 2018.

[40] S. Uddin, I. Haque, H. Lu, M. A. Moni, and E. Gide, "Comparative performance analysis of K-nearest neighbour (KNN) algorithm and its different variants for disease prediction," *Sci Rep*, vol. 12, no. 1, p. 6256, Apr. 2022, doi: 10.1038/s41598-022-10358-x.

[41] L. Auria and R. A. Moro, "Support Vector Machines (SVM) as a Technique for Solvency Analysis," *SSRN Electronic Journal*, 2008, doi: 10.2139/ssrn.1424949.

[42] A. Natekin and A. Knoll, "Gradient boosting machines, a tutorial," *Front Neurorobot*, vol. 7, 2013, doi: 10.3389/fnbot.2013.00021.

[43] A. Pindarwati and A. W. Wijayanto, "Measuring performance level of smart transportation system in big cities of Indonesia comparative study: Jakarta, Bandung, Medan, Surabaya, and Makassar",2015 International Conference on Information Technology Systems and Innovation (ICITSI), pp.1-6, 2015, IEEE

[44] S. R. Putri, A. W. Wijayanto, and S. Pramana, "Multi-source satellite imagery and point of interest data for poverty mapping in East Java, Indonesia: Machine learning and deep learning approaches", Remote Sensing Applications: Society and Environment,vol. 29, 100889, 2023, Elsevier

[45] Y. C. Putra, and A. W. Wijayanto, "Automatic detection and counting of oil palm trees using remote sensing and object-based deep learning", Remote Sensing Applications: Society and Environment,vol. 29, 100914, 2023, Elsevier