



Article

Identifying Key Indicators for Successful Foreign Direct Investment through Asymmetric Optimization Using Machine Learning

Aleksandar Kemiveš^{1,2}, Milan Randelović³ , Lidija Barjaktarović¹, Predrag Đikanović⁴, Milan Čabarkapa⁵ and Dragan Randelović^{4,*} 

¹ Department for Postgraduate Studies, The University Singidunum, 11000 Belgrade, Serbia; kemives@outlook.com (A.K.); lbarjaktarovic@singidunum.ac.rs (L.B.)

² PUC Infostan Technologies, 11000 Belgrade, Serbia

³ Science Technology Park, 18000 Niš, Serbia; milan.randjelovic@ntp.rs

⁴ Faculty of Diplomacy and Security, The University Union-Nikola Tesla Belgrade, 11000 Belgrade, Serbia; predragdjikanovic@fdb.edu.rs

⁵ Faculty of Engineering, The University of Kragujevac, 34000 Kragujevac, Serbia; mcabarkapa@kg.ac.rs

* Correspondence: dragan.randjelovic@fdb.edu.rs

Abstract: The advancement of technology has led humanity into the era of the information society, where information drives progress and knowledge is the most valuable resource. This era involves vast amounts of data, from which stored knowledge should be effectively extracted for use. In this context, machine learning is a growing trend used to address various challenges across different fields of human activity. This paper proposes an ensemble model that leverages multiple machine learning algorithms to determine the key factors for successful foreign direct investment, which simultaneously enables the prediction of this process using data from the World Bank, covering 60 countries. This innovative model, which adds to scientific and research knowledge, employs two sets of methods—binary regression and feature selection—combined in a stacking ensemble using a classification algorithm as the combiner to enable asymmetric optimization. The proposed predictive ensemble model has been tested in a case study using a dataset compiled from World Bank data across countries worldwide. The model demonstrates better performance than each of the individual algorithms integrated into it, which are considered state-of-the-art in these methodologies. Additionally, the findings highlight three key factors for foreign direct investment from the dataset, leading to the development of an optimized prediction formula.

Keywords: machine learning; binary regression; classification; feature selection; stacking ensemble method; prediction; indicators of successful foreign direct investments



Citation: Kemiveš, A.; Randelović, M.; Barjaktarović, L.; Đikanović, P.; Čabarkapa, M.; Randelović, D. Identifying Key Indicators for Successful Foreign Direct Investment through Asymmetric Optimization Using Machine Learning. *Symmetry* **2024**, *16*, 1346. <https://doi.org/10.3390/sym16101346>

Academic Editor: Zhixun Su

Received: 9 September 2024

Revised: 2 October 2024

Accepted: 9 October 2024

Published: 11 October 2024



Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

The background of the research presented in this paper focuses on identifying the most important predictors, a relevant and challenging issue across various fields, such as social life, medicine, and economics. This problem is also a significant research challenge. The primary motivation for the authors to conduct this study is the rapid economic growth in both developing and developed countries, where successful foreign direct investment (FDI) plays a crucial role in this development. FDI is the most common form of international capital movement, establishing long-term connections between the economies of different countries [1]. On the one hand, by investing directly in another country, companies from developed nations can avoid tariffs and other trade barriers while benefiting from lower production costs in developing countries. On the other hand, FDI greatly benefits host countries, primarily developing nations, by driving economic growth, creating jobs, and transferring knowledge. There are various types of FDI, but this research focuses on the

most common—FDI net inflow (measured as a percentage of gross domestic product (GDP)). This refers to the net inflows of investment aimed at acquiring a lasting management interest (10 percent or more of voting stock) in an enterprise that operates in a country different from that of the investor [2]. As noted in the literature [3], numerous factors influence FDI, such as GDP growth, official development assistance, trade, inflation, regulatory quality, government effectiveness, political stability, and population. Additionally, subgroups of these factors exist. It is essential to determine which of these many parameters are most significant in increasing FDI, allowing for targeted improvements based on the situation in countries worldwide.

To address the complex problem discussed in this paper, which involves conflicting requirements, various methodologies exist, such as multi-criteria decision-making methods and efficiency measurement techniques [4]. However, these methods are not the focus of this paper. Applying commonly used regression techniques to such problems is challenging due to the strict conditions required. For example, linear regression demands normality in the distribution of the sample, while binary regression requires a significantly larger number of samples than the number of factors considered [5]. In the context of this paper, where the percent of FDI in GDP for around 200 countries is treated as the dependent variable, the necessary condition is that for each independent variable (i.e., factor), there should be data from at least four or, preferably, more countries for each factor [5]. Recently, it has been recognized that machine learning (ML) classification techniques can be applied to this type of problem. In cases where two classes are involved, a two-class classifier can categorize the results into two groups, defined by a 2×2 confusion matrix, which can be used to predict significant factors without the limitations of regression techniques [6]. However, the literature still lacks sufficient references integrating multiple ML methods, whether different or of the same type, in ensemble prediction models. This gap motivated the authors to develop a novel approach. The need for further research in this area presents a significant challenge, one that the authors have embraced in this study. That is why, in the following paragraph, the current state of research in the field will be carefully reviewed, and key publications will be cited to provide context and support for the proposed methodology.

As mentioned at the beginning of this chapter, the literature review connected to the subject of this paper reveals a wide range of proposed methods that fall under multi-criteria decision-making (MCDM) and efficiency measurement (EM) approaches, which are used to determine significant factors for FDI. The MCDM group of methods can be divided into two categories: the first involves proposing novel methodologies of different types of MCDM for identifying the most significant indicators for FDI [7], while the second focuses on applying MCDM to solve problems in various fields across different countries worldwide [8]. Many studies, such as reference [8], could belong to both categories. The EM methods can also be classified into two subgroups: parametric [9] and non-parametric [10] methods. Since the methodologies employed in constructing the proposed model in this paper include regression analysis, feature selection, and supervised classification ML algorithms, the literature review emphasizes references that utilize these algorithms to solve the problem under consideration. One multiple regression model is discussed in [11] for determining significant FDI indicators in the Philippines. The authors of [12] explore the impact of FDI management on economic growth using a multiple linear regression model. Linhartova investigates FDI inflow in the Czech Republic using regression in [13]. In [14], Alharthi et al. investigate the macroeconomic and environmental factors attracting FDI inflows into Gulf Cooperation Council countries using a regression model. Al Mustofa et al. [15] analyze the impact of country security risks, legal and economic regulations, and different macroeconomic indicators on FDI inflows into 13 Muslim countries from 2002 to 2019. Researchers from City University in London [16] assess the determinants of FDI for OECD countries using a regression model and panel data covering 20 OECD countries from 1975 to 1997. Tamilselvan and Manikandan [17] investigate the relationship between FDI and GDP in India from 1991 to 2014, revealing that FDI positively impacts GDP. Colongeli [18] investigates the determinants of FDI inflows into Latin America and the

Caribbean using regression analysis. Chan et al. [19] use regression analysis to explore both short- and long-run flows of causality involving FDI, and in [20], regression models reveal that wage costs negatively affect FDI, while large regional markets, good infrastructure, and preferential policies have a positive impact on FDI in 29 Chinese regions. The authors of [21] evaluate the performance of five different artificial neural networks to understand the most important traits affecting safflower seed yield, and [22] studies the drivers of economic and financial integration across countries using random forest regression to overcome pitfalls of regression-based variable selection methods. A study in [23] compares the forecasting performance of neural networks and regression in forecasting economic series in developing economies. Alon et al. [24] propose a greenfield FDI attractiveness index using factor analysis and ML. The authors of reference [25] utilize several ML models, such as neural networks, K-Nearest Neighbors, trees, and support vector machines, to generate trading signals on various indices between 2002 and 2023. Jiménez and Herrero [26] use random forest regression to identify the features that yield the best results in internationalization for FDI in Spain. Various feature selection techniques applied to macroeconomic forecasting in Iran using World Bank Development Indicators are explored in [27], while [28] applies ML algorithms to select indicators attracting FDI to Hungary. The authors of [29] propose a methodology using Bagging and Naive Bayes for feature selection in classification problems with many features and few samples. A univariate feature selection algorithm based on the best-worst multi-attribute decision-making method is presented in [30]. Ensemble feature selection in binary ML classification is discussed in [31]. Reference [32] provides us with one comprehensive feature selection: a literature review. In [33], an adaptive hybrid-based ensemble method is proposed to improve binary classification performance, and [34] describes a hybrid approach for selecting and combining data mining models to construct ensemble models. The author of [35] constructs an ensemble effort estimation software to predict project effort using heterogeneous ensemble methods with filter feature selection, while [36] applies ant colony optimization to configure stacking ensembles for data mining. Finally, [37] reviews well-known ensemble techniques, focusing on Bagging, boosting, and stacking, while proposing a training set for improved prediction. Study [38] applies novel stacking ensemble methods to determine key atmospheric parameters for health incidents.

It is well known that solving the type of problem addressed in this paper depends on the chosen methodology and, in many cases, on the selected dataset. Therefore, the authors opted for the proposed stacking ensemble model because it improves accuracy and reduces overfitting from a methodological standpoint. It is also applicable and suitable for different types of predictors and addresses potential class imbalances within the dataset. This is what differentiates the proposed model from those discussed in the literature review. By using this approach, the authors developed a model that leverages the strengths of heterogeneous methodologies while mitigating their weaknesses.

Considering the mentioned premises, the main objective of this paper is to present research that discusses the advantages of aggregating two commonly used methodologies—traditional statistical binary regression and machine learning (ML) methods of feature selection [39]. These methodologies are combined into an ensemble optimization procedure to develop effective prediction models for identifying the most significant factors impacting FDI in countries worldwide. Since the classes of successful and unsuccessful countries in terms of FDI can be distinguished using a threshold value, which categorizes countries as dependent variables that fall into the successful class, the authors approached this issue as a classification task. To solve this problem, they applied a stacking ensemble model [40]. The proposed model, whose functional block diagram is presented in Figure 1, employs binary regression techniques for initial and final fine-tuning, as well as filter-based feature selection algorithms to rank the significance of each factor contributing to successful FDI in a country or region. A classification or other supervised ML algorithm is used as a combiner to assess the model's accuracy, enabling asymmetric optimization of the problem's dimensionality. The ultimate goal is to identify the most important factors

influencing FDI. Given the large number of different groups of indicators (i.e., factors), the authors chose to focus on one commonly defined group of factors—referred to as the “biggest obstacles” in World Bank terminology—applied to at least four countries per factor, based on World Bank data. The primary goal of this paper is to address the following research question: Is it possible to combine a feature selection machine learning (ML) method, which reduces the number of factors, with a traditional binary regression method in a stacking ensemble model using a classification algorithm as the combiner, to produce a model with better performance than either method individually? To evaluate the proposed model and validate this hypothesis, the authors applied it to a case study using data from the World Bank, covering various countries over a five-year period from 2017 to 2021. The case study examines the average net inflow of FDI data for 60 countries and evaluates a group of 15 factors, classified as the “biggest obstacle” factors, between 2017 and 2021. It seeks to determine the individual impact of the most significant factors on FDI during this period.

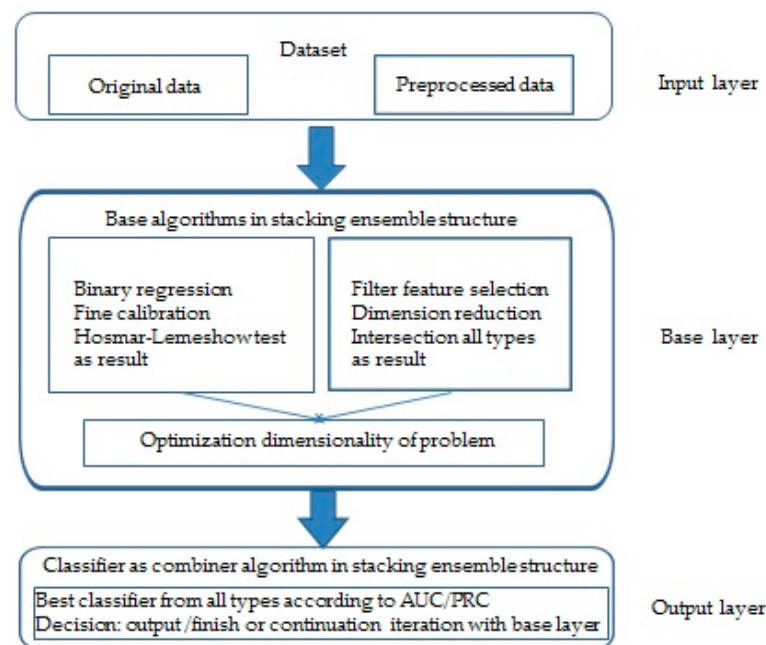


Figure 1. The functional block diagram of the proposed model.

The research described in this paper is expected to result in two main scientific and research contributions of the proposed model:

- **Methodological Contribution:** The novel model proposed in this research offers a new method for determining the significance of included factors in a univariate problem and making predictions using an ML-based ensemble algorithm. This model demonstrates superior characteristics compared to existing state-of-the-art methods integrated into it;
- **Technological Contribution:** The proposed model has practical implications, with the potential for real-world implementation as a useful web application.

To achieve the goal of developing an effective model for the problem under consideration, the rest of this manuscript is organized into the following sections: After the Introduction, which provides an overview of the research and the relevant literature, the “Materials and Methods” section describes the data and methodology used. The “Results” section presents and discusses the findings from the case study, identifying the most important factors and the proposed prediction model. Finally, the Conclusion provides a summary of the research, its contributions, and potential future work.

2. Materials and Methods

Considering the literature review on improved solutions for determining the significance of FDI indicators, many of which are computer-based and often utilize machine learning (ML) techniques, it can be concluded that ensemble ML methods are currently a trend for solving such complex problems. However, the existing literature still lacks a sufficient number of references that integrate multiple ML methods, whether different or of the same type. This gap motivated the authors to conduct further research into such methods.

To evaluate the proposed model, the authors used datasets available on the World Bank's website, which are detailed later in this section. For practical implementation, the dataset first required preprocessing, as described in a subsequent part of this paper. The preprocessed data were then analyzed, classifying all instances where the average FDI as a percentage of GDP was greater than 5% during the period from 2017 to 2021 into two categories: positive and negative (for values below 5%). This classification allowed the positive class to include countries where the conditions were favorable enough for the occurrence of successful FDI.

2.1. Methods

The problem under consideration, with the described dataset preprocessing, falls into the category of classification problems. Two main groups of methods are available for solving such problems: traditional statistical methods like logistic regression and feature selection.

In a logistic regression model, when the dependent variable takes on a finite set of values, the relationship between predictors—which can be continuous, binary, or categorical—and the dependent variable, which in this case is binary, is described. For binary outcomes, we implement binary regression, as is the case here. However, if the dependent variable has three or more categories, nominal logistic regression may be applied. Additionally, if the dependent variable has three or more categories that can be ranked, though the distances between them are not necessarily equal, ordinal logistic regression is appropriate. One might question whether linear regression can still be used in classification problems. The dependent variable is treated as a Bernoulli random variable, denoted as BinaryVariable in Equation (1) in the case of binary regression, where the two categories are coded as 0 or “false” for failure and 1 or “true” for success.

$$\text{BinaryVariable} = \begin{matrix} 0(\text{false}) - \text{failure} \\ 1(\text{true}) - \text{success} \end{matrix} \quad (1)$$

Since the dependent variable follows a Bernoulli distribution rather than being a continuous random variable, the errors cannot be normally distributed. Additionally, applying linear regression would result in nonsensical fitted values, possibly falling outside the {0, 1} range. In cases involving a binary dependent variable, one potential approach is to classify a “success” if the predicted value exceeds 0.5 and a “failure” if it does not. This method is somewhat comparable to linear discriminant analysis, which will be discussed later. However, this technique only produces binary classification results. When the predicted values are near 0.5, the confidence in the classification decreases. Furthermore, if the dependent variable contains more than two categories, this method becomes unsuitable, requiring the use of linear discriminant analysis instead.

Machine learning (ML) is a broad discipline grounded in statistical analysis and artificial intelligence, focused on acquiring knowledge, such as learning rules, concepts, and models that should be interpretable and accepted by humans. In the ML process, it is essential to validate the knowledge acquired, meaning the learned rules, concepts, or models must undergo evaluation. Two main evaluation methods exist, both involving the division of the available dataset into learning and testing sets in different ways.

The first method is the holdout test suite, where the dataset is split into two non-overlapping subsets: one for training and the other for testing the classifier (e.g., a

70:30 ratio). The classification model is built using the training data, and its performance is evaluated using the test data, allowing an assessment of classification accuracy based on the test results.

The second method, K-fold cross-validation, is more effective than using a single test set. In this approach, the dataset is split into k equal parts (or folds). One fold is used for testing, while the remaining folds are used for training. Predictions are made based on the current fold, and the process is repeated for k iterations, ensuring each fold is used for testing exactly once. The accuracy of the learned knowledge is one key measure of success, defined as the ratio of successful classifications to the total number of classifications. Other common evaluation metrics include precision, recall, the F1 measure, and, importantly, the Receiver Operating Characteristic (ROC) curve. Additionally, when dealing with imbalanced datasets, the precision–recall curve (PRC) may be more relevant, as will be discussed later in this chapter.

An important fact that must be noted is that the variables choice in the classification process affects all performance metrics of classification. Therefore, different techniques for variable selection are necessary during the data preparation or preprocessing phase, and dimension reduction methods may also be applied.

This paper proposes an ensemble ML model for determining the significance of various FDI indicators and predicting the likelihood of successful FDI in a given country based on these factors. As mentioned earlier, the proposed method integrates several feature selection algorithms, a binary regression algorithm, and the best-performing classification algorithm, combining them into a stacking ensemble ML method. The following subsections will briefly describe the methodologies employed, as the ensemble method integrates logistic binary regression with ML-based classification methods and feature selection techniques.

2.1.1. Classification Methodology

Classification algorithms are part of supervised machine learning (ML) and are commonly used for predictive modeling tasks. In the novel model proposed in this paper, these algorithms are utilized as the combiner in a stacking ensemble method. At the start of the procedure, the best algorithm is selected from several types across different groups, ideally choosing the best option from each group. The classification methodology relies on the existence of labeled instances in more than one class (or attribute) of objects, allowing it to classify the categorical class (attribute) value based on the remaining factors or attributes [41]. Selecting the appropriate classification algorithm for a specific application is not only the first step but also one of the most critical aspects of the ML process, which is especially important when such methodology is applied to large datasets. To address the problem considered in this paper, the proposed ensemble model uses a classification approach that categorizes instances into two classes: positive or negative, corresponding to “true” or “false”. The possible outcomes of this classification are displayed in the confusion matrix, as shown in Table 1.

Table 1. The confusion matrix for the binary two-class classification.

		Predicted Label	
		Positive	Negative
Actual label	Positive	TP (true positive)	FN (false negative)
	Negative	FP (false positive)	TN (true negative)

Let N denote the total number of members in the considered set, as shown in Table 1. This value is the sum of positive and negative cases, i.e., $TP + FN + FP + TN = N$, where TP represents true positives, FN false negatives, FP false positives, and TN true negatives. All results presented in Table 1, for the case of two-class classification, can be used to

calculate the most important classification metrics—accuracy, precision, recall, and the F1 measure—using the following formulas:

$$\text{Accuracy} = (\text{TP} + \text{TN}) / \text{N} \in [0, 1] \quad (2)$$

$$\text{Precision} = \text{TP} / (\text{TP} + \text{FP}) \in [0, 1] \quad (3)$$

$$\text{Recall} = \text{TP} / (\text{TP} + \text{FN}) \in [0, 1] \quad (4)$$

$$\text{F1 measure} = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}} \in [0, 1] \quad (5)$$

In evaluating the performance of any classifier, the Receiver Operating Characteristic (ROC) curve is commonly used as one of the most important measures. It represents the false positive rate on the OX axis and the true positive rate on the OY axis [42]. For instance, point (0, 1) signifies perfect classification, where all samples are correctly classified, while point (1, 0) indicates a classification where all samples are incorrectly classified. The output in ROC space generated by naive neural networks or Bayes classification algorithms is a score—a numeric value that represents probability—whereas discrete classifiers yield a single point. In both cases, they express the likelihood that a particular instance belongs to a specific class [43]. The area under the curve (AUC) is a commonly used metric for measuring the accuracy of a model, with AUC values greater than 50% considered acceptable and values above 70% indicating good classification performance.

For imbalanced datasets, however, the precision–recall curve (PRC) is a more suitable measure than the ROC AUC [44]. Similar to the ROC curve, PRC plots are generated by connecting pairs of precision and recall values at each threshold. A good classifier will have a PRC that approaches the upper-right corner. In general, the closer a point is to the position where both precision and recall are 100%, the better the model's performance. Like with ROC, the area under the precision–recall curve (AUPRC) is a reasonable measure of performance, with AUPRC > 0.5 indicating acceptable performance and higher values reflecting better classifier performance.

In practical terms, classification is a machine learning task but can also be considered a data mining task involving the separation of instances in a dataset into one of the predetermined classes based on the values of input variables [45]. The literature review shows that the most commonly applied classifiers include Bayes networks, decision trees, neural networks, and K-Nearest Neighbor, among others. For the proposed model, it is essential to use at least five of the most popular classification algorithms. The authors selected algorithms from different groups as categorized in WEKA software, version 3.8.6 [46], including types from Bayes, Meta, Trees, Misc, Rules, Lazy, and Functions. Below, a brief description of a selected algorithm from each of these groups is provided.

The Naive Bayes classifier [47] is one of the oldest classification algorithms and generates a model based on Bayes' theorem. The term “naive” refers to the simplifying assumption it makes: the factors used in classification are conditionally independent, and there are no hidden factors that could influence the classification. These assumptions allow the Naive Bayes classifier to perform classification efficiently. For conditionally independent factors A_1, A_2, \dots, A_k , the probability of the class factor A is calculated using the following rule:

$$P(A_1, \dots, A_k | A) = \prod_{i=1}^k (A_i | A) \quad (6)$$

The main advantage of this classifier is the convenience of small datasets.

Bagging, or Bootstrap Aggregating, is an ensemble method from the Meta group of classifiers that enhances the stability and accuracy of weak estimators or classification models. Breiman [48] introduced Bagging as a technique for reducing variance in a given base model, such as decision trees or other methods that involve selecting variables and fitting them into a linear model. Random forest [49] belongs to the Trees group of commonly used

classifiers. Tree-structured classifiers are an attractive choice for solving one classification or prediction problem because they are easy to interpret.

The PART classifier [50] from the Rules group in Weka builds partial decision trees. Each iteration utilizes the C4.5 decision tree algorithm to generate the best leaf and derive a corresponding rule for the tree. This approach is particularly useful in binary classification, as applied in this paper.

SMO [51], from the Functions group in Weka, is an efficient optimization algorithm used in the implementation of support vector machines (SVMs). Although it is not one of the most commonly used classifiers, it is applied in this paper due to its suitability for binary classification with both numerical and binary factors, which aligns with the problem being addressed.

InputMappedClassifier belongs to the Misc group of Weka, but the IBk classifier belongs to the Lazy group, and both of them are not oft-used classifiers, and that is why we will not waste space describing them in more detail [52].

2.1.2. Logistic Regression

In logistic regression, when solving a problem using machine learning (ML) methodology, it is important to use probabilistic classifiers that not only return the label for the most likely class but also provide the probability of that class. These probabilistic classifiers can be evaluated using a calibration plot, which shows how well the classifier performs on a given dataset with known outcomes—this is especially relevant for the binary classifiers considered in this paper. For multi-class classifiers, separate calibration plots are required for each class.

In the proposed model, the authors applied the basic idea of univariate calibration, where logistic regression transforms classifier scores into probabilities of class membership in a two-class scenario. Many other authors have extended this concept to multi-class cases, as in [53].

The primary objective of logistic regression is to produce the best-fitting model that explains the relationship between a dichotomous dependent variable (the characteristic of interest) and a set of independent variables. Logistic regression generates coefficients in a formula that predicts or classifies a logit transformation of the probability of the characteristic's presence, often denoted as p (including the standard error and significance level). This is defined as the logged odds, represented by $\text{logit}(p)$:

$$\text{logit}(p) = b_0 + b_1X_1 + b_2X_2 + \dots + b_kX_k \quad (7)$$

$$\text{odds} = \frac{p}{1-p} = \frac{\text{probabilityofcharacteristicspresence}}{\text{probabilityofcharacteristicsabsence}} \quad (8)$$

$$\text{logit}(p) = \ln\left(\frac{p}{1-p}\right) \quad (9)$$

The coefficients in the logistic regression equation are represented by $b_0, b_1, b_2, \dots, b_k$. These coefficients indicate whether the corresponding independent variables have an increasing or decreasing effect on the dependent variable, with $b_i > 0$ indicating an increasing effect and $b_i < 0$ indicating a decreasing effect. When the independent variables are dichotomous, their impact on the dependent variable can be determined by simply comparing their regression coefficients. By exponentiation of both sides of the regression equation (as shown in Equations (7) and (9)), the equation can be transformed into a well-known form of logistic regression:

$$\text{odds} = \frac{p}{1-p} = e^{b_0} \cdot e^{b_1X_1} \cdot e^{b_2X_2} \cdot e^{b_sX_s} \cdot \dots \cdot e^{b_kX_k} \quad (10)$$

As is evident from the provided Formula (10), when variable X_i increases by 1 unit and all other parameters remain unchanged, then the odds will increase by a value of parameter e^{b_i} .

$$e^{b_t(1+X_t)} - e^{b_t X_t} = e^{b_t X_t} = e^{b_t(1+X_t)-b_t X_t} = e^{b_t+b_t X_t-b_t X_t} = e^{b_t} \quad (11)$$

The factor e^{b_i} represents the odds ratio (O.R.) for the independent variable X_i .

It indicates the relative change in the odds of the outcome: when the O.R. is greater than 1, the odds increase, and when it is less than 1, the odds decrease. This change occurs when the value of the independent variable increases by one unit.

Logistic regression can be implemented using various statistical software programs, with SPSS [54] being one of the most commonly used tools. SPSS provides three basic methods for binary regression: the enter method, the stepwise method, and the hierarchical method. In this paper, the authors employed the standard enter method for the proposed model. In the hierarchical method, researchers determine the order in which independent variables are added to the model. Stepwise methods include two categories: forward selection and backward elimination. The basic characteristic of the enter method is that it includes all independent variables in the model simultaneously. All methods aim to remove independent variables that are weakly correlated with the dependent variable from the regression equation.

2.1.3. Future Selection Techniques

Machine learning classification methods are sensitive to data dimensionality, making it evident that applying various dimensionality reduction techniques can significantly improve results. Algorithms for feature subset selection perform a space search based on candidate evaluation [55]. Several evaluation measures have proven effective in removing irrelevant and redundant features, including the consistency and the correlation measures. The consistency measure seeks to identify the minimum number of features that consistently differentiate class labels, defining inconsistency as instances where two cases have different class labels but share the same feature values.

The most common taxonomy for feature selection methods divides them into three groups [56]:

- Filter: Known examples include Relief, GainRatio, and InfoGain;
- Wrapper: Notable examples include BestFirst, GeneticSearch, and RankSearch;
- Embedded: These methods combine filter and wrapper techniques.

Weka, a widely used, free-to-use software, includes a feature selection function that reduces the number of attributes by applying different algorithms. This made it the tool of choice for evaluating the proposed model in the case study representing the problem discussed in this paper.

Since the first group of filter-based feature selection methods was used in the proposed model, the algorithms from this group are briefly described below. For a dataset denoted as S , the filter algorithm begins by creating an initial subset D_1 , which could be an empty set, the entire set, or a randomly selected subset. It then explores the feature space based on a predetermined search strategy. Each subset D_i generated during the search is evaluated using an independent metric and compared to the current best subset. If it performs better, it becomes the new best subset. The search continues until a predefined stopping condition is met. The final output of the algorithm is the last best subset, which is considered the final result. The feature selection process often relies on entropy as a metric for assessing the purity of a set of examples, considering the measure of unpredictability in the system. The entropy of Y is as follows:

$$H(Y) = - \sum_{y \in Y} p(y) \cdot \log_2(p(y)) \quad (12)$$

Feature selection methods vary in how they handle the issues of irrelevant and redundant attributes. In the proposed model, the authors utilized multiple filter algorithms, more than the recommended minimum of five, covering all the filter algorithms available in the WEKA 3.8.6 software. All these algorithms were used with the Ranker search method, which produces a ranked list of attributes based on their individual evaluations. This method must be paired with a single-attribute evaluator, not an attribute-subset evaluator. In addition to ranking attributes, Ranker also selects them by eliminating those with lower rankings.

Considering that entropy can serve as a criterion for impurity in training set S , there is a way to define a measure that reflects the additional information each attribute provides, as determined by the class. This measure, known as information gain, represents the amount by which the entropy of the attribute is reduced. It is denoted as InfoGain and is used to evaluate the value of an attribute in relation to the class, calculated using the following formula:

$$\text{InfoGain}(\text{Class}, \text{Attribute}) = H(\text{Class}) - H(\text{Class}|\text{Attribute}) \quad (13)$$

where H represents the entropy of information, and the information gained about an attribute after observing the class is equal to the information gained when the observation is reversed.

The information gain ratio, referred to as GainRatio, is an asymmetrical measure that corrects the bias inherent in the InfoGain measure. It is essentially a modified version of InfoGain, designed to reduce its bias toward certain attributes, and is calculated using the following formula:

$$\text{GainRatio} = \frac{\text{InfoGain}}{H(\text{Class})} \quad (14)$$

As shown in Formula (13), when predicting a specific variable or attribute, InfoGain is normalized by dividing it by the entropy of the class and vice versa. This normalization ensures that the GainRatio values fall within the range $[0, 1]$. A GainRatio of 1 means that knowledge of the class perfectly predicts the variable or attribute, while a GainRatio of 0 indicates no relationship between the variable or attribute and the class.

FilteredAttributeEval (ClassifierAttributeEval) is a classifier that handles nominal and binary classifications with various types of attributes, including nominal, string, relational, binary, unary, and even missing values. It uses an arbitrary evaluator on data processed through a filter built solely from the training data.

SymmetricalUncertAttributeEval, described in Equation (15), is a classifier that evaluates the significance of an attribute by calculating its symmetrical uncertainty, considering the presence of each class in the process.

$$\text{SymmU}(\text{Class}, \text{Attribute}) = 2 * (H(\text{Class}) - H(\text{Class}|\text{Attribute})) / (H(\text{Class}) + H(\text{Attribute})) \quad (15)$$

This classifier handles nominal, binary, and missing class classifications using attributes such as nominal, binary, unary, and others.

ReliefFAttributeEval is an instance-based classifier that randomly samples instances and examines neighboring instances from both the same and different classes, handling both discrete and continuous class data.

PrincipalComponents transforms the attribute set by ranking the new attributes according to their eigenvalues. A subset of attributes can optionally be selected by choosing enough eigenvectors to represent a specified portion of the variance, with the default set to 95%.

CorrelationAttributeEval evaluates the importance of an attribute by measuring its correlation (Pearson's) with the class. For nominal attributes, each value is treated as an indicator and considered on a value-by-value basis. The overall correlation for a nominal attribute is calculated as a weighted average.

2.1.4. Ensemble Methods ML

As mentioned earlier in the Introduction and at the start of this section, ensemble methods are based on the concept that combining algorithms of different types can yield better results than each algorithm individually. There are several types of ensemble methods and their taxonomies, with the most commonly used being the following:

- Bootstrap Aggregating (Bagging);
- Boosting;
- Stacking.

The following is a summary, as found in the literature [57]:

- Ensemble learning combines multiple machine learning algorithms into a single model to enhance performance. Bagging primarily aims to reduce variance, boosting focuses on reducing bias, and stacking seeks to improve prediction accuracy;
- While ensemble methods generally offer better classification and prediction results, they require more computational resources than evaluating a single model within the ensemble. Thus, ensemble learning can be seen as compensating for less effective learning algorithms by performing additional computations. However, it is important to note that in many problems, including the case study presented in this paper, real-time computation is not a constraint, making this extra computational effort manageable.

Stacking

The stacking ensemble algorithm involves training multiple machine learning algorithms and combining their predictions or classifications into a single model. This approach typically yields better performance than any individual algorithm alone [58]. Stacking can be applied to both supervised learning tasks and unsupervised learning tasks. In stacking, each algorithm is trained using the available data, and then a meta-algorithm is trained to make the final estimation, classification, or prediction. This process often involves cross-validation to prevent overfitting [59]. While logistic regression is commonly used as the combiner algorithm in practice, the proposed model in this article uses feature selection algorithms for this role.

2.1.5. Proposed Ensemble Model

The authors acknowledge the potential negative effects of poor model fit in regression and the impact of imbalanced data in feature selection and classification, which has been noted in the prior literature [60,61]. As a result, they question the reliance on regression or feature selection combined with classification as the primary methods for solving binary classification problems, such as the one they are addressing in this paper. Therefore, as mentioned in the Introduction, the authors chose to apply a stacking methodology that incorporates both binary regression and feature selection methods, with a suitable classification algorithm serving as the combiner in the proposed model.

The proposed stacking ensemble method integrates two types of machine learning algorithms in an asymmetric structure. The first is a binary regression method that initiates the model and evaluates the goodness of fit at each step, while the second is a feature selection algorithm that reduces the dimensionality of the problem by selecting fewer factors. This process continues as long as the classification algorithm, acting as the combiner, permits it. The combiner only grants permission if the dimensionally reduced problem yields better values in terms of PRC (for imbalanced datasets) or ROC AUC (for balanced datasets) before the iterative process begins.

The goodness of the regression model is assessed using the Hosmer–Lemeshow test, and ultimately, the significance coefficients of the final regression model with acceptable fit will refine the prediction by identifying the most important factors.

In this way, the authors aim to develop an optimized iterative procedure that combines the strengths of both methods while minimizing their weaknesses. Although both binary regression and feature selection algorithms are widely recognized as supervised learning techniques used for predictions on labeled datasets, their divergent approaches to binary classification and other machine learning problems highlight their differences.

The proposed model is provided with the algorithm presented in Figure 2 and described in Algorithm 1.

Algorithm 1: Determining the importance of indicators for successful FDI

```

1. * Input data each instace with n1 factors for m instances-countries and preprocess the data.
NEXT
    Perform binary regression and determine  $n \leq n1$  non-colinear input indicators;
    Check regressions goodness  $HLsig \geq 0.05$ 
IF NO No valid prediction GOTO END
ELSE
NEXT
    Check datasets imbalance OneClass is  $\leq 25\%$  of OtherClass
IF NO No Treshold TR = ROC AUC
ELSE Treshold TR = PRC
NEXT
2. ** Perform classification with a minimum of five classification algorithms of different types and
identify the algorithm 'TheBest' with the highest PRC (or ROC AUC) value.
    Check goodness of classification  $PRC|(ROC AUC)| \geq 0.7$ 
IF NO No valid prediction GOTO END
ELSE
NEXT
3. *** Apply feature selection procedure using minimum 5 different filter algorithms; Using
intrsection logic operation determine  $M \leq N$  attributes;
NEXT
4. **** Using TheBest classification algorithm determine with dataset of M attributes new its value
TheNewBest
NEXT
    Check regressions goodness  $HLsig \geq 0.5$ 
IF NO GOTO 6 *****
ELSE GOTO 5 *****
NEXT
5. **** Check goodness of classification  $PRC|(ROC AUC)| \geq 0.7$ 
IF NO GOTO 6 *****
ELSE GOTO 3 ***
NEXT
6. ***** By means of already carried out in the previous Step 5 binary regression determine
important indicators for FDI i.e., prediction formula
END

```

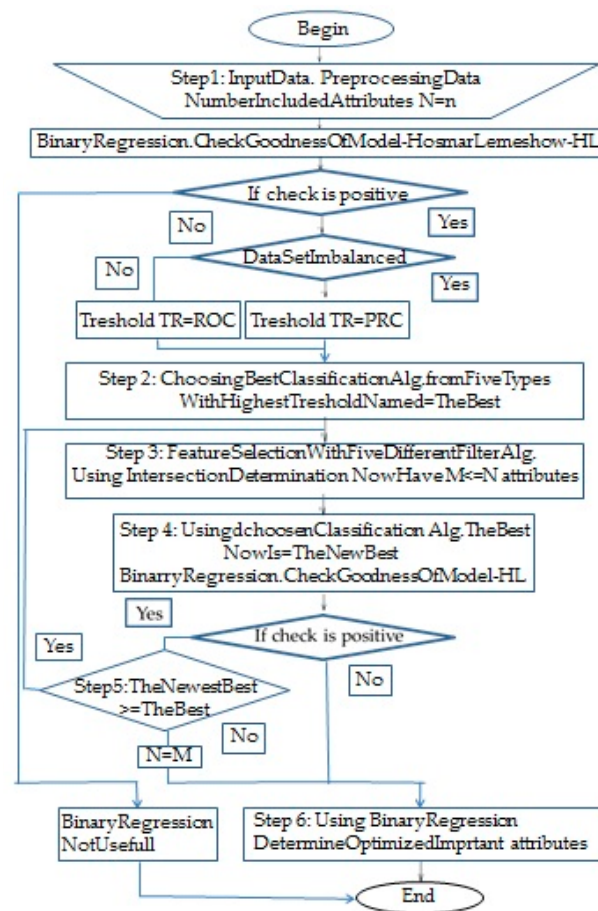


Figure 2. The flowchart of Algorithm 1.

* Step 1. Step 1 starts with an obligatory preprocessing dataset with n_1 indicators and m instances—countries as independent variables. The last special column contains binary values of the dependent variable FDI in % of GDP, which indicates the success of FDI in each specific country. To be useful for the application preprocessed dataset, it must have a minimum of four instances, i.e., regions or countries for each used indicator of FDI and successful instances classified as true, each of which has an inflow of FDI in GDP greater or equal to five percent. Before entering the iterative loop, binary regression is made, and collinearity is checked to exclude potentially collinear indicators. The goodness of the Hosmer–Lemeshow model is tested, and if it is greater than 0.05, the procedure can start; otherwise, it is not applicable and cannot determine the significance of individual indicators. Also, before entering the procedure in Step 2, determination of the imbalance of the considered dataset is performed, and in the case of imbalance, i.e., if one of the two classes in the considered binary classification is present with less than 25% in the classification procedure, the PRC evaluator is used as the most significant in the selection of the best from a minimum of 5 different types of classification algorithms; otherwise, the ROC measure is used.

** Step 2. This step involves selecting the best classification algorithm for the model from at least five different types of classification algorithms. It concludes by evaluating the goodness of classification. If the value of the most significant measure, PRC (or ROC AUC), is less than 0.7, the procedure terminates, as it would not be applicable and cannot determine the significance of individual indicators. A 10-fold cross-validation test procedure is used.

*** Step 3. The loop itself begins with Step 3, in which a potential reduction in the dimensionality of the problem is determined between a minimum of five feature selection

filter algorithms using a logical function of the intersection of individually obtained results, and the algorithm continues in Step 4.

**** Step 4. In Step 4, with a reduced number of indicators selected in Step 3 by the best classification algorithm, the PRC is determined by comparing an unbalanced set or ROC in the opposite case. It first examines the goodness of the regression model for that reduced number of indicators. If it is OK, it moves on to Step 5, and if there is no fulfillment of this condition, the procedure ends with a previously determined number of indicators in Step 6.

***** Step 5. In this step, it is checked whether the value of PRC, i.e., ROC, is now greater than or equal to the previous one. In the case of fulfillment of this condition, the loop continues with Step 3, and in the opposite case, the loop is exited in Step 6.

***** Step 6. If the optimization procedure for a specific dataset is possible, this algorithm ends with a previously determined number of indicators in Step 6, where significant indicators are determined based on the value of the regression model, and a prediction model can also be provided. The next step leads to the end of this algorithm.

2.2. Materials

The proposed model for estimating weight coefficients in this study is based on data and reports from the World Bank's Enterprise Analysis Unit of the Development Economics Global Indicators Department [62]. However, the original datasets from the World Bank required preparation to be suitable for addressing the problem discussed in this paper. As a result, the data had to be preprocessed for use in the first step of the proposed model.

2.2.1. Dataset World Bank for FDI Countries around the World

The World Bank Enterprise Surveys (WBESs), part of the World Bank Enterprise Analysis Unit within the Development Economics Global Indicators Department, offer a vast array of economic data covering over 219,000 firms across 159 economies, with the expectation of reaching 180 economies soon. These surveys provide valuable insights into various aspects of the business environment, such as firm performance, access to finance, infrastructure, and more. The data are publicly available and are particularly useful for scientists, researchers, policymakers, and others. The data portal offers access to over 350 WBESs, 12 Informal Sector Enterprise Surveys in 38 cities, Micro-Enterprise Surveys, and other cross-economy databases.

The Enterprise Surveys focus on factors influencing the business environment, which can either support or hinder firms' operations. A favorable business environment encourages firms to operate efficiently, fostering innovation and increased productivity, both of which are crucial for sustainable development. A more productive private sector leads to job creation and generates tax revenue essential for public investment in health, education, and other services. Conversely, a poor business environment presents obstacles that impede business activities, reducing a country's potential for growth in terms of employment, production, and overall welfare.

These surveys, conducted by the World Bank and its partners, cover all geographic regions and include small, medium, and large firms. The surveys are administered to a representative sample of firms in the non-agricultural formal private economy. The survey universe, or population, is uniformly defined across all countries and includes the manufacturing, services, transportation, and construction sectors. However, sectors such as public utilities, government services, health care, and financial services are excluded. Since 2006, most Enterprise Surveys have been implemented under a global methodology that includes a uniform universe, uniform implementation methodology, and a core questionnaire.

The Enterprise Surveys collect a wide range of qualitative and quantitative data through face-to-face interviews with firm managers and owners, focusing on the business environment and firm productivity. The topics covered in the surveys are grouped into 13 categories, comprising over 100 indicators that impact FDI. These categories include firm characteristics, gender, workforce, performance, innovation and technology, infrastruc-

ture, trade, finance, crime, informality, regulations and taxes, corruption, and the biggest obstacles to doing business [63]. From the World Bank—Data Bank World Development Indicators [64], the authors have taken the data, which show what percent of FDI is in GDP for each country the authors include in this study.

2.2.2. Preprocessed Dataset World Bank for FDI 60 Countries around the World

The prepared and processed dataset that the authors used to evaluate the proposed model is provided as a Supplementary File for this paper. It is obtained on the following premises of the authors.

Keeping in mind the original data of the World Bank, the authors noticed the necessary reprocessing of the same in the next steps:

1. For more than a hundred indicators provided in 13 groups, a correct analysis would require about 500 instances, i.e., countries, and there are not that many in the world;
2. Data exist separately for companies of different sizes, but there are also aggregated data;
3. For individual countries, data for indicators as independent variables in the research are collected at intervals of about 5 years, and data for the dependent variable in the research for the percentage of FDI in the GDP of an individual country are available annually.

For these reasons, the authors, in the preprocessing of the dataset usable for the intended research, took data for a sufficient number of 60 countries that exist in the period of 5 years, 2017–2021. The independent variable included aggregate data for companies of all sizes and the average investment percentage in the same period for each of the countries included in the study. The dependent variable was shown to be successful in terms of FDI for the percentage of FDI greater than 5.

3. Results

To evaluate the impact of selected socio-economic factors influencing the success of FDI on the GDP of any country worldwide, the authors analyzed data from 60 countries, utilizing 15 indicators identified by the World Bank as the biggest obstacles to doing business in its open-access dataset. These indicators include access to finance, access to land, business licenses and permits, corruption, courts, crime, theft and disorder, customs and trade regulations, electricity, inadequately educated workforce, labor regulations, political instability, informal sector practices, tax administration, tax rates, and transportation. The case study is based on the data acquired from the World Bank Enterprise Analysis Unit of the Development Economics Global Indicators Department over the last several tens of years, but the authors singled out from that dataset as sufficient, and according to the way in which the data are collected, for the current moment, a five-year period of time from 2017 to 2021.

Data analysis was performed using two methodologies, binary regression and filter feature selection, organized in one ensemble ML model with a third classification as a combiner, as is already described in Section 2.1.5. The proposed novel model holds six steps as separate entities of execution of the tasks foreseen by the algorithm, and because of that, this section is divided into the next six, Sections 3.1–3.6, to enable a clear description of applied methodology in each step of the proposed procedure as well as better presentation and understanding of obtained results. It will offer a concise and clear presentation of the experimental results, their interpretation, and a discussion of these results.

3.1. Input of Preprocessed Data for the Considered Case Study—Initial Binary Regression

In Step 1, mandatory preprocessing of the data available on the aforementioned link of the World Bank website is carried out with n_1 general—in our case, 15 indicators—and m instances (countries) general—in our case, 60 of them—as independent variables. The last special column contains binary values (true or false) of dependent-variable FDI in % of GDP, which indicates the success of FDI in a particular country and has the value true only

if the inflow of FDI in GDP is greater or equal to five percent. We apply binary regression to the entered prepared data in order to first exclude from the data those indicators that are collinear, and we do not find collinear indicators. At the same time, we examine the goodness of the regression part of the model, and we perform the Hosmer–Lemeshow test. Because it is greater than 0.05, the procedure can continue; otherwise, it could not be applicable to determine the significance of individual indicators and prediction, and the procedure should finish at this first step. At the end of this step, the imbalance of the used dataset is checked, and because we have only 11 from a total of 60 true, i.e., success-characterized countries ($11/60 = 18.3\% < 25\%$), the PRC evaluator will be used as the most significant in the determination of the best form. In the next step, we use seven different types of classification algorithms.

The results obtained with binary regression applied on preprocessed data are listed in Table 2, in which the meaning of the abbreviations used for measures in the columns are described:

B—Denotes the unstandardized regression weight;

S.E.—Measures how much the unstandardized regression weight can vary. It is similar to a standard deviation to a mean;

Wald—Denotes the test statistic for the individual predictor variable like multiple linear regression has a *t* test, logistic regression has a χ^2 test, and it determines the Sig. value;

df—This is the number of degrees of freedom for the model. There is one degree of freedom for each predictor in the model;

Sig.—Determines significant variables, and a *p* value below 0.050 is considered significant;

Exp(B) or Odds Ratio—Denotes the OR ratio that represents the change in the probability of belonging to one outcome category when the value of the predictor increases by one measurement unit;

95% C.I. (Confidentiality Interval) OR—Represents the 95% C.I. for the odds ratio, which means that with these values, we are 95% certain that the true value of the odds ratio is between those units. But, if the C.I. contains a value of 1 between the given borders of the interval, the odds ratio will not be significant.

Binary logistic regression analysis was used to examine the correlation between a success percent FDI in GDP as a dependent variable, on the one hand, and all indicators from the group of the biggest obstacle as independent variables, which are access to finance, access to land, business licenses and permits, corruption, courts, crime, theft and disorder, customs and trade regulations, electricity, inadequately educated workforce, labor regulations, political instability, practices of the informal sector, tax administration, tax rates, and transportation, keeping in mind that in our case study, the higher values for each indicator individual should be understood as a better-rated case for this indicator, but it is not realistic that this is the case for a successful percent of FDI in GDP as a dependent variable.

The calculated values in the column marked ‘Sig.’ are all less than 0.05, indicating that these indicators are not significant in this model, and all positive data in column B have as a consequence that increasing values in such column will increase the probability of a successful percent of FDI in GDP as the dependent variable. Also, the value 1 included in the confidential interval C.I. for all indicators shows that the odds ratio is not significant, i.e., this wide range indicates that the model does not produce precise predictions. The obtained results confirm that *N* = 15 factors are valid but not significant for prediction.

Those conclusions confirm a logical and experiential already-known expectation but show all mentioned deficiencies of binary regression in solving the binary classification problem presented in the considered case study. Further data and discussion about the results of necessary tests in this binary regression for the continuation of the proposed model are provided in the tables below.

Table 2. OR values and their 95% CI for assessing the impact of the examined 15 indicators for FDI.

	B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I. for EXP(B)	
							Lower	Upper
Access to finance	10.559	6.740	2.45	1	0.117	38,509.23	0.071	2.10×10^{10}
Access to land	10.506	6.684	2.47	1	0.116	36,547.14	0.075	1.78×10^{10}
Business licenses and permits	10.334	6.711	2.37	1	0.124	30,746.01	0.060	1.58×10^{10}
Corruption	10.541	6.734	2.45	1	0.117	37,848.43	0.070	2.04×10^{10}
Courts	10.404	6.609	2.47	1	0.115	32,992.59	0.078	1.39×10^{10}
Crime, theft and disorder	10.897	6.832	2.54	1	0.111	54,027.46	0.083	3.53×10^{10}
Customs and trade regulations	11.398	6.896	2.73	1	0.098	89,186.05	0.120	6.61×10^{10}
Electricity	10.556	6.719	2.46	1	0.116	38,392.07	0.073	2.01×10^{10}
Inadequately educated workforce	10.678	6.762	2.49	1	0.114	43,387.969	0.076	2.47×10^{10}
Labor regulations	10.453	6.762	2.39	1	0.122	34,654.613	0.061	1.97×10^{10}
Political instability	10.730	6.776	2.50	1	0.113	45,701.205	0.078	2.67×10^{10}
Practices of the informal sector	10.835	6.822	2.52	1	0.112	50,754.932	0.079	3.25×10^{10}
Tax administration	9.862	6.515	2.29	1	0.130	19,190.865	0.055	6.74×10^9
Tax rates	10.741	6.792	2.50	1	0.114	46,211.982	0.077	2.79×10^{10}
Transportation	11.332	6.905	2.69	1	0.101	83,482.687	0.111	6.29×10^{10}
Constant	-1071.49	677.119	2.50	1	0.114	0.000		

The obtained results provide a summary of the good performances of the proposed model, beginning from the so-called Omnibus test, which is a goodness-of-fit test with a value for significance of 0.03, showing that our regression model has a good prediction of results. Then, the classification test provided in Table 3, with a value of 90% correct prediction, has good classification, and, of course, in the end, the most important indicator for the goodness of the model is the results of the Hosmer–Lemeshow test, 0.729, which enables the continuation of this algorithm to Step 2.

Table 3. Classification table.

		Classification Table ^a		Predicted	
		Observed	FDI in % of GDP False	True	Percentage Correct
Step 1	FDI in % of GDP	FALSE	48	1	98.0
		TRUE	5	6	54.5
Overall Percentage					90.0

^a. The cut value is 0.500.

3.2. Selection of the Classification Algorithm—Combiner in Proposed Stacking Ensemble

In Step 2 of this algorithm, selection is executed for the best classification algorithm for this model from seven algorithms with different types of classification as it is in WEKA (Table 4) and which are already described in Section 2.1.1 of this paper: Bayes, Naive Bayes, Meta, Bagging, Trees, Random Forest, Rules, Part, Function, SMO, Lazy, IBk, Misc, InputMappedClassifier. Ten cross-validation tests were used.

Table 4. Selection of the best classification algorithm on the used dataset.

	Precision	Recall	F1 Measure	Accuracy	ROC AUC	PRC
Bayes.Naive Bayes	0.694	0.683	0.689	68.333	0.427	0.689
Meta.Bagging	0.768	0.817	0.761	81.666	0.630	0.780
Trees.Random Forest	0.768	0.817	0.761	81.666	0.572	0.736
Rules.Part	0.650	0.717	0.682	71.666	0.494	0.701
Function.SMO	/	0.817	/	81.666	0.500	0.701
Lazy.IBk	0.647	0.700	0.673	70.000	0.429	0.684
Misc.InputMappedClassifier	/	0.817	/	81.666	0.450	0.688

It is evident from Table 4 that the Bagging algorithm, from the Meta classification algorithms, has the best characteristics, i.e., the best values for each classification measure, including the most important PRC, which value we assign to a variable, i.e., TheBest = 780.

Step 2 ends with checking the goodness of classification, and because the value of the most significant measure PRC is bigger than 0.7, the procedure continues with Step 3 of this procedure.

3.3. Filter Feature Selection Algorithm as the Second Classification of Proposed Stacking Ensemble

The iterative loop itself starts in Step 3 with a potential reduction in the dimensionality of the problem and using seven different filter feature selection algorithms, which are already described in Section 2.1.3 of this paper: GainRatio, InfoGain, FilteredAttributeEval (ClassifierAttributeEval), SymmetricalUncertAttributeEval, ReliefFAttributeEval, PrincipalComponents, and CorrelationAttributeEval.

By applying the logical function of the intersection of the results obtained individually from each filter algorithm, we determined that the number of indicators could be reduced from 15 to 13. This reduction was made because PrincipalComponents selected 13 indicators, while the other algorithms identified all 15 as significant.

The 13 selected indicators are access to finance, access to land, business licenses and permits, corruption, courts, crime, theft and disorder, customs and trade regulations, electricity, inadequately educated workforce, labor regulations, political instability, informal sector practices, and tax administration. At the end of this step, the algorithm continues on to Step 4.

3.4. Checking Goodness of Regression Model with Reduced Indicators

In Step 4, with a reduced number of 13 indicators selected in Step 3 by the best Bagging classification algorithm, we determine the new value of PRC = 0.807 and assign this value to the variable TheNewBest = 0.807. After that, we examine the goodness of the regression model for that reduced number of indicators, and because the Hosmer–Lemeshow test on the prepared data shows a significance of $0.709 > 0.050$, which is a good value, we continue on to Step 5.

3.5. Checking Value Most Important Measure of Classification for Possible Iteration Continuation

In Step 5 of the proposed procedure, we check whether the value of PRC TheNewBest determined in Step 4 is greater than the previous value remembered as TheBest, and because $0.807 > 0.780$, we continue the iteration loop again with Step 3. In the opposite case, the loop must be exited in Step 6, which directly leads to the end of the procedure.

In the second pass through the iterative loop, and in each subsequent pass, we begin again with Step 3 as previously described, performing feature selection once more. In this pass, we determine, in the same manner, that 12 indicators could be considered significant. These are access to finance, access to land, business licenses and permits, corruption, courts, crime, theft and disorder, customs and trade regulations, electricity, inadequately educated workforce, labor regulations, political instability, and informal sector practices.

On these reduced data with 12 indicators, we apply the classification Bagging algorithm and determine that we now have a value for PRC = 0.811 and assign this value to the TheNewBest variable. We also implement on these data binary regression, from which we determine using the Hosmer–Lemeshow test that we still have a good significance of $0.490 > 0.05$ and also how we can see from data provided in Table 5's variable equation that for a dataset with 12 indicators, we now have significant indicators, such as customs and trade regulations, and also negative correlation indicators, such as business licenses and permits and labor regulations, with successful FDI percent in GDP as a dependent variable, which is a more realistic expected case, and because of that, better binary classification. So we check again in Step 5 if TheNewBest is greater than the previous, and because $0.811 > 0.807$, we try to continue this iterative procedure of optimization for solving one binary classification problem and continue with Step 3, which is at the beginning of the loop.

Table 5. OR values and their 95% CI for assessing the impact of the examined 12 indicators for FDI.

		Variables in the Equation					95% C.I. for EXP(B)		
		B	S.E.	Wald	df	Sig.	Exp(B)	Lower	Upper
Step 1 ^a	Accesstofinance	−0.076	0.068	1.260	1	0.262	0.927	0.812	1.058
	Accesstoland	0.300	0.214	1.957	1	0.162	1.350	0.887	2.054
	Businesslicensesandpermits	−0.181	0.223	0.658	1	0.417	0.835	0.539	1.292
	Corruption	−0.076	0.138	0.302	1	0.583	0.927	0.707	1.215
	Courts	0.058	0.270	0.046	1	0.831	1.060	0.624	1.800
	Crimetheftanddisorder	0.053	0.180	0.085	1	0.771	1.054	0.740	1.501
	Customsandtraderegulations	0.388	0.179	4.705	1	0.030	1.474	1.038	2.094
	Electricity	−0.075	0.105	0.511	1	0.475	0.928	0.755	1.140
	Inadequatelyeducatedworkforce	0.041	0.052	0.636	1	0.425	1.042	0.942	1.153
	Laborregulations	−0.242	0.166	2.125	1	0.145	0.785	0.567	1.087
	Politicalinstability	0.005	0.060	0.006	1	0.939	1.005	0.894	1.129
	Practicesoftheinformalsector	0.082	0.060	1.838	1	0.175	1.085	0.964	1.221
Constant	−2.379	2.959	0.647	1	0.421	0.093			

^a. Variable(s) entered on Step 1: Accesstofinance, Accesstoland, Businesslicensesandpermits, Corruption, Courts, Crimetheftanddisorder, Customsandtraderegulations, Electricity, Inadequatelyeducatedworkforce, Laborregulations, Politicalinstability, Practicesoftheinformalsector.

In the third pass through the iterative loop, we start again with Step 3 and perform the described feature selection again. In this pass, we similarly determined that 12 indicators could be considered significant. These indicators are access to finance, access to land, business licenses and permits, corruption, courts, crime, theft and disorder, customs and trade regulations, electricity, inadequately educated workforce, labor regulations, and political instability.

On these reduced data with 11 indicators, we apply the classification Bagging algorithm and determine that we now have a value for PRC = 0.793 and assign this value to the TheNewBest variable and implement on these data binary regression, from which we determine using the Hosmer–Lemeshow test that we still have good significance. But because in Step 5 the checked TheNewBest is now less than the previous, i.e., $0.793 < 0.811$, we must go on to Step 6 to continue this iterative procedure and, from this step, exit the proposed procedure.

3.6. Checking Value Most Important Measure of Classification for Possible Iteration Continuation

In Step 6, which is the last step before the proposed procedure ends, this optimization procedure for specific datasets in each of the three passes through the iterative loop, which is made possible by Steps 3–6, and the proposed procedure ends with a last determined number of 12 indicators, with which the set conditions are met. The significant indicators are determined based on the value of the last binary regression model, which is used for fine calibration of the end result. It shows exactly one indicator as dominant significant—Customsandtraderegulations—and three more—Laborregulations, Accesstoland, and Practicesoftheinformalsector—that are close to the limit for them to be so. A prediction model based on the last results shown in Table 5 can be provided if there is a need. The next step of the proposed procedure leads to the end of this algorithm.

4. Discussion

4.1. Discussion of Obtained Result for Most Important Indicators for Successful FDI and Possible Prediction

As presented in previous subsections, the case study considered in this paper used the proposed stacking ensemble ML model for evaluation. The authors employed seven different types of filter feature selection algorithms (such as GainRatio and Relief) for dimensionality reduction, binary logistic regression for evaluation and fine calibration within the model, and Bagging as the combiner algorithm, chosen as the best classifier from seven different types of classification algorithms (including LogitBoost, J48 decision tree, and Naive Bayes).

The results showed that the proposed algorithm optimizes the asymmetric procedure for determining the significance of the 15 World Bank-selected “Biggest Obstacle” indicators on the successful inflow of FDI as a percentage of GDP across 60 countries worldwide. This optimization was achieved through dimensionality reduction using feature selection, fine calibration via binary logistic regression, and the classification algorithm as the combiner. The results also indicated that this process led to a unique prediction formula with strong classification characteristics, qualifying the proposed ensemble method as superior to each individual method included in the aggregation. It also outperformed other state-of-the-art ensemble methods, such as Random Forest, AdaBoost, and Bagging, as shown in Table 6.

Table 6. Comparison performance indicators using proposed and known ensemble procedures.

	Precision	Recall	F1 Measure	Accuracy	ROC AUC	PRC
Proposed model	/	0.817	/	81.666	0.714	0.811
Random forest	0.768	0.817	0.761	81.666	0.572	0.736
Ada Boost	0.721	0.750	0.734	75.000	0.442	0.695
Bagging	0.768	0.817	0.761	81.666	0.630	0.780

It should be noted that the obtained results in the proposed model use one evaluation using 10 cross-validation methods that are state-of-the-art for this mandatory process [65].

The initial basic hypothesis comprehended in this paper in its introductory part was that it is possible to construct one ensemble model of ML that tests a successful percent of FDI in GDP in some countries around the world and has better characteristics than those included in this ensemble and other known algorithms that are state-of-the-art for solving such type of so-called binary classification problems. The results of the applied analyses show that it is possible and confirm this hypothesis, and in this way, it is enabled to determine the most important indicators and one prediction formula. Namely, analyzing the data from Table 5, we can conclude that the successful percent of FDI in GDP (SPF-GDP) dominantly depends on the set of Customsandtraderegulations (CDR), then from Labourregulations (LR), then from Accesstoland (AL), then from Practicesoftheinformalsector (PIS), then from Accesstofinance (AF), and so on. If we include in this set each indicator with a max significance ≤ 0.2752 , all of them have a positive direction for the dependent variable

except Accesstofinance, but only indicator Customsandtraderegulations has a significance of $0.030 < 0.050$, and it indicates significance (OR = 1.474 95% CI: 1.038 to 2.094; $p = 0.030$). Dependent variables do not depend significantly on the rest of the considered indicators.

The prediction formula could be provided as it is shown in Equation (16):

$$\text{SPF} - \text{GDP} = -2.379 + 0.388\text{CDR} - 0.242\text{LR} + 0.3\text{AL} + 0.082\text{PIS} - 0.076\text{AF} \quad (16)$$

4.2. Discussion of Proposed Methodology

The proposed stacking ensemble model enables data compression through dimensionality reduction, thereby reducing storage requirements, and eliminates the problem of overfitting and using unbalanced datasets. Additionally, as a stacking ensemble algorithm, it has the following effects:

- Increases stability and helps eliminate redundant features, if present;
- Reduces variance;
- Minimizes noise in the dataset.

All of these characteristics lead to an increase in PRC or AUC in the case of balanced datasets and an increase in accuracy without compromising other commonly used measures of goodness in binary classification problems. “Asymmetry in ensemble modeling” refers to the unequal contribution of two key methods in the model. Specifically, the greater contribution comes from feature selection, as it reduces the dimensionality of the problem, which in turn reduces errors and improves stability. This process consequently optimizes accuracy and other performance metrics essential for effective binary classification.

As mentioned in the previous subsection, all results from the classification algorithms included in the proposed model were obtained using the 10-fold cross-validation technique. This was employed to address the issues of overfitting and underfitting while simultaneously enhancing the generalization capabilities of the proposed model by testing it on both training data and test data.

It is important to note the following limitations of the proposed method, which generally apply to all ensemble methodologies:

- The longer execution time, as one component of computational complexity, is greater than the time required for any individual algorithm in the ensemble. However, this limitation does not affect the real-time application of the model for the considered problem;
- The lack of diversity in the ensemble when models are trained independently or sequentially without considering their interactions. This limitation, however, is mitigated by the use of a 10-fold cross-validation method, which is state-of-the-art for such processes;
- The improvement of the ensemble model over the best individual model tends to be relatively small, especially when the base algorithms are already sophisticated. However, this is not the case in our research.

As for the computational complexity of each machine learning model used in the proposed method, the authors believe this topic is outside the scope of the current paper and may be explored in future work.

In computing, the computational complexity of an algorithm refers to the total resources required to run it, typically in terms of time and memory. As mentioned in the Limitations section, the longer execution time of the proposed algorithm is not a barrier to its application in the considered and other non-real-time problems. The space complexity of the proposed algorithm is also not an obstacle, as it is compensated for by the dimensionality reduction in the problem. For these reasons, we did not delve into detailed calculations of these complexities in this paper, but this could be addressed in future work if such characteristics become the focus of new research.

The authors plan the future work in this field in two directions:

- Functional—With consideration of the same problem n , the different grouped regions for this but and other types of problems in human life;
- Methodological—Inclusion of other different forms of ensemble models in solving such types of binary classification problems.

From a practical and functional standpoint, the authors plan to continue their research in this field first of all by including other groups of indicators for successful percent of FDI in GDP using more countries around the world than the now-included 60 and with partial consideration of that problem by regions around the world. Also, the authors will use the proposed methodology described in this paper to determine the most important indicators for successful percent of FDI in GDP in other problems in several different problems from other fields of human life, such as medicine, health, traffic, economy, education, etc., which the authors have already done [66].

On the other side, they plan to expand the research of the proposed methodology by the inclusion of a bigger number of classifications as well as filter ranking algorithms as a part of the proposed ensemble model and other types of machine learning algorithms in the application of ensemble models, including the application of ablation study as a special technique.

The authors carefully considered the inclusion of an ablation study in this proposed model and concluded that it could be highly beneficial for future work, particularly in ensemble methods aimed at solving binary classification problems, especially when using a neural network as the combiner. However, for the current solution, an ablation study is not appropriate, as the model is based on the premise that a specific number of classification algorithms are necessary to optimize dimensionality, which in turn enhances performance as measured by AUC and PRC. This approach answers the research question posed in this study.

5. Conclusions

The primary goal of this paper, defined in the Introduction, was to address the following research question: Is it possible to combine a feature selection machine learning (ML) method, which reduces the number of factors, with a traditional binary regression method in a stacking ensemble model using a classification algorithm as the combiner, and thereby produce a model with better characteristics than either method individually? The research conducted in this paper provides a positive answer. The ensemble model proposed in this manuscript represents an asymmetrically optimized procedure based on the stacking algorithm of ensemble ML, incorporating logistic regression and feature selection algorithms while using a classification algorithm as the combiner. Employing feature selection facilitates dimensionality reduction in the binary classification problem of determining the importance of various indicators from the “biggest obstacle” group on the successful percentage of FDI in GDP across sixty countries worldwide. The results demonstrated that the proposed algorithm outperforms each individual method in the ensemble, as well as other state-of-the-art ensemble algorithms in ML, in determining the significance of selected indicators for successful FDI participation in GDP. Additionally, it produces a unique prediction formula. The research yielded two significant outcomes:

- From a scientific perspective, the authors have proposed a new iterative asymmetric optimization procedure with excellent binary classification performance, which can be applied to both prediction and discriminative classification problems. This approach can be used to assess the importance of individual attributes in multivariate problems;
- Practically, the proposed algorithm, as part of the iterative algorithm family, can be applied to similar binary classification problems in various fields of human activity.
- For future research, the authors propose two main areas of focus;
- First, they aim to improve the dataset by including additional groups of indicators for successful FDI participation in GDP and expanding the study to more countries, dividing them by regions worldwide. Additionally, the authors plan to apply the proposed

- methodology to identify the most important indicators for successful FDI participation in GDP, as well as to address different problems in other fields of human life;
- Second, they consider enhancing the proposed methodology by incorporating a larger number of classification algorithms and filter ranking algorithms into the ensemble model and other types of machine learning algorithms in the application of ensemble models, including the application of ablation study as a special technique.

Supplementary Materials: The following supporting information can be downloaded at: <https://www.mdpi.com/article/10.3390/sym16101346/s1>.

Author Contributions: Conceptualization, A.K., M.R., L.B., P.Đ., and D.R.; methodology, D.R. and M.R.; software, M.Č.; validation, D.R.; formal analysis, A.K. and M.R.; investigation, A.K., M.R., L.B., and P.Đ.; resources, A.K. and L.B.; data curation, P.Đ.; writing—original draft preparation, D.R.; writing—review and editing, M.R. and M.Č.; visualization, A.K., L.B., and D.R.; supervision, D.R.; project administration, A.K., M.R., L.B., and P.Đ.; funding acquisition, L.B. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: Data are contained within the article and Supplementary Materials.

Acknowledgments: We thank the Enterprise Analysis Unit of the Development Economics Global Indicators Department of the World Bank for the data. Also, we thank the Scientific Technological Park in Niš, Serbia, for all its support.

Conflicts of Interest: Author Aleksandar Kemiveš is employed by the PUC Infostan Technologies. Author Milan Randelović is employed by the Science Technology Park Niš. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest

References

1. Rahmonov, T.; Kalimbetov, X.K.; Baymuratova, Z.A.; Sarsenbayev, B.A. Foreign Direct Investment: Importance for the Development of the Country's Economy. *Solid State Technol.* **2020**, *63*, 9804–9812.
2. Zeqiri, N.; Bajrami, H. Foreign Direct Investment (FDI) Types and Theories: The Significance of Human Capital. In Proceedings of the International Conferences on Business, Technology and Innovation 2016, Durres, Albania, 28–30 October 2016; pp. 43–58. [\[CrossRef\]](#)
3. Blonigen, B.A. A review of the empirical literature on FDI determinants. *Atl. Econ. J.* **2005**, *33*, 383–403. [\[CrossRef\]](#)
4. Nguyen, P.H.; Nguyen, T.-L.; Le, H.-Q.; Pham, T.-Q.; Nguyen, H.-A.; Pham, C.-V. How Does the Competitiveness Index Promote Foreign Direct Investment at the Provincial Level in Vietnam? An Integrated Grey Delphi–DEA Model Approach. *Mathematics* **2023**, *11*, 1500. [\[CrossRef\]](#)
5. Narain, H.; Mahesh, B.; Mukherjee, S. Impact of Macroeconomic Variables on FDI: Regression Analysis and Forecasting using Time Series Data. *HRC J. Econ. Financ.* **2023**, *1*, 57–84.
6. Singh, D. Foreign direct investment and local Interpretable model-agnostic Explanations: A rational framework for FDI decision making. *J. Econ. Financ. Adm. Sci.* **2024**, *29*, 98–120. [\[CrossRef\]](#)
7. Gupta, S.; Jha, B.; Singh, R. Decision making framework for foreign direct investment: Analytic hierarchy process and weighted aggregated sum product assessment integrated approach. *J. Public Aff.* **2021**, *22*, e2771. [\[CrossRef\]](#)
8. Randelović, M.; Nedeljković, S.; Jovanović, M.; Čabarkapa, M.; Stojanović, V.; Aleksić, A.; Randelović, D. Use of Determination of the Importance of Criteria in Business-Friendly Certification of Cities as Sustainable Local Economic Development Planning Tool. *Symmetry* **2020**, *12*, 425. [\[CrossRef\]](#)
9. Zulkarnain, R. Stochastic Frontier Model Incorporating Spatial Effect to Measure Efficiency Component of Multifactor Productivity. In Proceedings of the Asia-Pacific Statistics Week, Bangkok, Thailand, 15–19 June 2020.
10. Lei, M.; Zhao, X.; Deng, H.; Tan, K.-C. DEA analysis of FDI attractiveness for sustainable development: Evidence from Chinese provinces. *Decis. Support Syst.* **2013**, *56*, 406–418. [\[CrossRef\]](#)
11. Sabroso, L.; Cañete, A. Factors Driving Foreign Direct Investment: An Empirical Investigation Using Multiple Regression. *J. Asian Dev.* **2023**, *9*, 16–31. [\[CrossRef\]](#)
12. Matušovičová, M.; Matušovičová, S. The Impact of Foreign Direct Investment Management on Economic Growth Using Multiple Linear Regression (MLR). *TEM J.* **2023**, *12*, 2326–2332. [\[CrossRef\]](#)
13. Linhartová, V. Factors influencing the foreign direct investment inflow in the Czech republic. *Acta Acad. Karviniensia* **2018**, *18*, 36–46. [\[CrossRef\]](#)
14. Alharthi, M.; Islam, M.; Alamoudi, H.; Murad, W. Determinants that attract and discourage foreign direct investment in GCC countries: Do macroeconomic and environmental factors matter? *PLoS ONE* **2024**, *15*, e0298129. [\[CrossRef\]](#) [\[PubMed\]](#)

15. Al Mustofa, M.U.; Sukmana, R.; Herianingrum, S.; Ratnasari, R.T.; Mawardi, I.; Zulaikha, S. Determining Factors of Inward Foreign Direct Investment (FDI) in Selected Muslim Countries. *J. Econ. Coop. Dev.* **2021**, *42*, 1–26. Available online: <https://jeed.sesric.org/pdf.php?file=ART20062602-2.pdf> (accessed on 1 August 2024).
16. Agiomirgianakis, G.; Asteriou, D.; Papathoma, K. The Determinants of Foreign Direct Investment: A Panel Data Study for the OECD Countries. City University London–Department of Economics, School of Social Sciences. Discussion Paper Series No.03/06. Available online: https://www.city.ac.uk/_data/assets/pdf_file/0019/90424/0306_agiomirgianakis-et-al.pdf (accessed on 1 August 2024).
17. Tamilselvan, M.; Manikandan, S. A Study on Impact of Foreign Direct Investment on Gross Domestic Production in India. *Int. J. Acad. Res. Bus. Soc. Sci.* **2015**, *5*, 224–233. Available online: <https://mpr.ub.uni-muenchen.de/73349/> (accessed on 10 October 2024). [[CrossRef](#)] [[PubMed](#)]
18. Colongeli, N. The Determinants of Foreign Direct Investment: Evidence from Latin America and the Caribbean. Available online: <https://digitalcommons.bryant.edu/cgi/viewcontent.cgi?article=1044&context=eeb> (accessed on 15 July 2024).
19. Chan, M.W.L.; Hou, K.; Li, X.; Mountain, D.C. Foreign direct investment and its determinants: A regional panel causality analysis. *Q. Rev. Econ. Financ.* **2014**, *54*, 579–589. [[CrossRef](#)]
20. Cheng, L.K.; Kwan, Y.K. What are the determinants of the location of foreign direct investment? The Chinese experience. *J. Int. Econ.* **2000**, *51*, 379–400. [[CrossRef](#)]
21. Abdipour, M.; Kwan, Y.K. Artificial neural networks and multiple linear regression as potential methods for modeling seed yield of safflower (*Carthamus tinctorius* L.). *Ind. Crops Prod.* **2018**, *127*, 185–194. [[CrossRef](#)]
22. Akbari, A.; Ng, L.; Solnik, B. Drivers of economic and financial integration: A machine learning approach. *J. Empir. Financ.* **2021**, *61*, 82–102. [[CrossRef](#)]
23. Chuku, C.; Simpasa, A.; Oduor, J. Intelligent forecasting of economic growth for developing economies. *Int. Econ.* **2019**, *159*, 74–93. [[CrossRef](#)]
24. Alon, I.; Bretas, V.P.; Sclip, A.; Paltrinieri, A. Greenfield FDI attractiveness index: A machine learning approach. *Compet. Rev.* **2022**, *32*, 85–108. [[CrossRef](#)]
25. Grudniewicz, J.; Ślepaczuk, R. Application of machine learning in algorithmic investment strategies on global stock markets. *Res. Int. Bus. Financ.* **2023**, *66*, 023. [[CrossRef](#)]
26. Jiménez, A.; Herrero, Á. Selecting features that drive internationalization of Spanish firms. *Cybern. Syst.* **2019**, *50*, 25–39. [[CrossRef](#)]
27. Goldani, M. Evaluating Feature Selection Methods for Macro-Economic Forecasting, Applied for Iran’s Inflation Indicator. *arXiv* **2004**, arXiv:2406.03742. [[CrossRef](#)]
28. Singh, D.; Turala, M. Machine Learning and Regularization Technique to Determine Foreign Direct Investment in Hungarian Counties. *Danube* **2022**, *13*, 269–291. [[CrossRef](#)]
29. Nnamoko, N.; Arshad, F.; England, D.; Vora, J.; Norman, J. Evaluation of filter and wrapper methods for feature selection in supervised machine learning. *Age* **2014**, *21*, 33–2.
30. Abellana, D.P.M.; Lao, D.M. A New univariate feature selection algorithm based on the best–worst multi-attribute decision-making method. *Decis. Anal. J.* **2023**, *7*, 100240. [[CrossRef](#)]
31. Abellana, D.P.M.; Roxas, R.R.; Lao, D.M.; Mayol, P.E.; Lee, S. Ensemble feature selection in binary machine learning classification: A novel application of the evaluation based on distance from average solution (EDAS) method. *Math. Probl. Eng.* **2022**, *1*, 4126536. [[CrossRef](#)]
32. Kumar, V.; Minz, S. Feature selection: A literature review. *SmartCR* **2014**, *4*, 211–229. [[CrossRef](#)]
33. Fan, X.; Lung, C.; Ajila, S. Using Hybrid and Diversity-Based Adaptive Ensemble Method for Binary Classification. *Int. J. Intell. Sci.* **2018**, *8*, 43–74. [[CrossRef](#)]
34. Zhu, D. A hybrid approach for efficient ensembles. *Decis. Support Syst.* **2010**, *48*, 480–487. [[CrossRef](#)]
35. Hosni, M.; Idri, A.; Abran, A. Improved Effort Estimation of Heterogeneous Ensembles using Filter Feature Selection. In Proceedings of the 13th International Conference on Software Technologies, Porto, Portugal, 26–28 July 2018; Volume 1, pp. 405–412, ISBN 978-989-758-320-9. [[CrossRef](#)]
36. Chen, Y.; Wong, M.L.; Li, H. Applying Ant Colony Optimization to configuring stacking ensembles for data mining. *Expert Syst. Appl.* **2014**, *41*, 2688–2702. [[CrossRef](#)]
37. Loughrey, J.; Bi, Y.; Wu, S.; Nugent, C. A survey of commonly used ensemble-based classification techniques. *Knowl. Eng. Rev.* **2013**, *29*, 551–581. [[CrossRef](#)]
38. Randelovic, D.; Randelovic, M.; Cabarkapa, M. Using Machine Learning in the Prediction of the Influence of Atmospheric Parameters on Health. *Mathematics* **2022**, *10*, 3043. [[CrossRef](#)]
39. Korgaonkar, C. Analysis of the impact of financial development on foreign direct investment: A data mining approach. *J. Econ. Sustain. Dev.* **2012**, *3*, 70–79.
40. Kemiveš, A.; Barjaktarović, L.; Randelović, M.; Čabarkapa, M.; Randelović, D. Assessing the Efficiency of Foreign Investment in a Certification Procedure Using an Ensemble Machine Learning Model. *Mathematics* **2024**, *12*, 1020. [[CrossRef](#)]
41. Romero, C.; Ventura, S.; Espejo, P.; Hervás, C. Data mining algorithms to classify students. In Proceedings of the 1st IC on Educational Data Mining (EDM08), Montreal, QC, Canada, 20–21 June 2008; pp. 20–21.
42. Vuk, M.; Curk, T. ROC curve, lift chart and calibration plot. *Metod. Zv.* **2006**, *3*, 89–108. [[CrossRef](#)]

43. Witten, H.; Eibe, F. *Data Mining: Practical Machine Learning Tools and Techniques*, 2nd ed.; Morgan Kaufmann Series in Data Management Systems; Elsevier: Amsterdam, The Netherlands, 2005.
44. Rainio, O.; Teuvo, J.; Klén, R. Evaluation metrics and statistical tests for machine learning. *Sci. Rep.* **2024**, *14*, 6086. [[CrossRef](#)]
45. Benoit, G. Data Mining. *Annu. Rev. Inf. Sci. Technol.* **2002**, *36*, 265–310. [[CrossRef](#)]
46. Weka (University of Waikato: New Zealand). Available online: https://waikato.github.io/weka-wiki/downloading_weka/ (accessed on 10 October 2024).
47. Berrar, D. Bayes' Theorem and Naive Bayes Classifier. *Encycl. Bioinform. Comput. Biol.* **2018**, *1*, 403–412. [[CrossRef](#)]
48. Breiman, L. Bagging predictors. *Mach. Learn.* **1996**, *24*, 123–140. [[CrossRef](#)]
49. Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5–32. [[CrossRef](#)]
50. Pham, T.M.; Chen, P.; Nguyen, T.; Yoon, S.; Bui, T.; Nguyen, A. PEEB: Part-based Image Classifiers with an Explainable and Editable Language Bottleneck. *arXiv* **2024**, arXiv:2403.05297v3. [[CrossRef](#)]
51. Keerthi, S.S.; Shevade, S.K.; Bhattacharyya, C.; Murthy, K.R.K. Improvements to Platt's SMO Algorithm for SVM classifier design. *Neural Comput.* **1999**, *13*, 637–649. [[CrossRef](#)]
52. Chahal, D. Comprehensive Analysis of Data Mining Classifiers using WEKA. *Int. J. Adv. Comput. Res.* **2018**, *9*, 718–723. [[CrossRef](#)]
53. Bella, A.; Ferri, C.; Hernández-Orallo, J.; Ramírez-Quintana, M.J. Calibration of machine learning models. In *Handbook of Research on Machine Learning Applications*; IGI Global: Hershey, PA, USA, 2009.
54. SPSS Statistics 17.0 Brief Guide. Available online: http://www.sussex.ac.uk/its/pdfs/SPSS_Statistics_Brief_Guide_17.0.pdf (accessed on 20 August 2024).
55. Liu, H.; Motoda, H. *Feature Selection for Knowledge Discovery and Data Mining*; Kluwer Academic: Boston, MA, USA, 1998.
56. Novaković, J. Rešavanje klasifikacionih problema mašinskog učenja. In *Bussines Process Reengineering*; Faculty of Technical Sciences Čačak, University of Kragujevac: Kragujevac, Serbia, 2013; Volume 4.
57. Available online: <https://scikit-learn.org/stable/modules/ensemble.html#stacking> (accessed on 20 August 2024).
58. Wolpert, D. Stacked generalization. *Neural Netw.* **1992**, *5*, 241–259. [[CrossRef](#)]
59. Zhou, Z.H. *Ensemble Methods Foundations and Algorithm*; Chapman and Hall/CRC: New York, NY, USA, 2012. [[CrossRef](#)]
60. Harrell, F. Hosmer-Lemeshow vs. AIC for Logistic Regression. Available online: <https://stats.stackexchange.com/q/18772> (accessed on 20 August 2023).
61. Terra, J. Regression vs. Classification in Machine Learning for Beginners. Available online: <https://www.simplilearn.com/regression-vs-classification-in-machine-learning-article> (accessed on 15 August 2024).
62. World Bank Enterprise Surveys. Available online: www.enterprisesurveys.org (accessed on 1 August 2024).
63. Available online: <https://www.enterprisesurveys.org/en/data> (accessed on 1 August 2024).
64. Available online: <https://databank.worldbank.org/reports.aspx?source=2&series=BX.KLT.DINV.CD.WD&country=> (accessed on 1 August 2024).
65. Arlot, S.; Celisse, A. A survey of cross-validation procedures for model selection. *Stat. Surv.* **2010**, *4*, 40–79. [[CrossRef](#)]
66. Mišić, J.; Kemiveš, A.; Randelović, M.; Randelović, D. An Asymmetric Ensemble Method for Determining the Importance of Individual Factors of a Univariate Problem. *Symmetry* **2023**, *15*, 2050. [[CrossRef](#)]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.