



RESEARCH ARTICLE

OPEN ACCESS

COMPARATIVE EVALUATION OF MACHINE LEARNING MODELS FOR PREDICTIVE MAINTENANCE USING THE AI4I 2020 DATASET: EFFECTS OF DATA BALANCING AND FEATURE SELECTION

Juan Carlos Gonzalez Islas¹, Jesus Vargas-Ortiz², Aldo Márquez-Grajales³,
Ernesto Bolaños-Rodríguez⁴, Asdrúbal López-Chau⁵, Rosa María Ortega-Mendoza⁶

^{1, 2, 3, 4, 6}Universidad Autonoma Del Estado De Hidalgo. Mexico.

⁵Universidad Autonoma Del Estado De Mexico. Mexico.

¹<https://orcid.org/0000-0002-2190-0660>, ²<https://orcid.org/0009-0006-6546-8319>, ³<https://orcid.org/0000-0001-9567-2069>,
⁴<https://orcid.org/0000-0002-1432-7720>, ⁵<https://orcid.org/0000-0001-5254-0939>, ⁶<https://orcid.org/0000-0003-4506-9653>

E-mail: juan_gonzalez7024@uaeh.edu.mx, va421350@uaeh.edu.mx, aldo_marquez@uaeh.edu.mx, ebolanos@uaeh.edu.mx,
alchau@uaemex.mx, rosa_ortega@uaeh.edu.mx

ARTICLE INFO

Article History

Received on: March 03, 2026

Reviewed on: April 02, 2026

Accepted on: May 01, 2026

Published on: July 31, 2026

Keywords:

Machine Learning,
Predictive Model,
Feature Selection,
Imbalanced Data,
Industrial Failures.

ABSTRACT

This article presents a comparative analysis of machine learning (ML) models applied to machine fault classification, using the AI4I 2020 dataset. The performance of five algorithms (random forest, decision tree (J48), artificial neural networks (ANN), k-nearest neighbors (KNN), and Naive Bayes) is evaluated in six different data scenarios: synthetic minority oversampling (SMOTE) technique, undersampling, and original imbalanced distribution in the complete feature set and in the feature selection set. In addition, the impact of feature selection techniques on predictive power is evaluated. The goal is to identify the optimal predictive model and compare the performance of these approaches to determine the advantages of feature reduction and data balancing strategies. To ensure robust validation, the models are analyzed using precision, recall, specificity, accuracy, and F1 score. The main conclusions obtained demonstrate that it is possible to build reliable models capable of maximizing fault detection and minimizing false alarms, which is vital for predictive maintenance systems in industries where accuracy is critical.



Copyright ©2026 by authors and Galileo Institute of Technology and Education of the Amazon (ITEGAM). This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

1. INTRODUCTION

Industrial maintenance has undergone a transformation with the advancement of Industry 4.0 principles, evolving from predominantly reactive approaches to Predictive Maintenance (PdM) strategies. This approach is based on the analysis of real-time operational data to anticipate the occurrence of failures, optimize asset life, and reduce operating costs associated with unplanned maintenance [1]. In this scenario, Machine Learning plays a fundamental role by enabling the modeling of complex machinery behaviors, overcoming the limitations of traditional statistical methods [2]. However, the effectiveness of these models depends critically on the quality and structure of the available data. Industrial environments simultaneously present two significant problems: class imbalance and data scarcity, due to the low frequency of failure events, and high dimensionality associated with the use of multiple sensors, which can introduce noise and redundant information. Recent literature suggests that addressing these problems in isolation is insufficient, and that it is necessary to analyze the interaction between dimensionality reduction techniques and strategies for subsampling unbalanced data [3] [4]. This study uses the AI4I 2020 [5] Predictive Maintenance dataset, available in the UCI repository, which is continuously used as a reference in predictive maintenance research. Several studies have explored solutions to class imbalance and high dimensionality in predictive maintenance in order to reduce the bias toward normal machine operation. For instance, Sridhar and Sanagavarapu [6] showed that using SMOTE is an effective strategy for handling data imbalance in industrial environments, as it allows models to learn failure patterns more accurately.

Likewise, recent work by Bezerra et al. [7] analyzed how feature selection directly affects the predictive performance of algorithms, while authors such as Cioch, Kulisz, and Gola [8] validated the effectiveness of comparative models to establish performance benchmarks in industrial applications. The relevance of this interaction lies in the fact that the application of advanced classification algorithms does not guarantee optimal performance when data presents structural biases or is contaminated by irrelevant variables, in addition to presenting class imbalance. In particular, ignoring these characteristics can lead to an increase in false negatives, i.e., situations in which critical failures are not detected despite obtaining high overall accuracy values. This scenario represents a significant risk to the continuity and operational safety of industrial systems. Therefore, there is a need to empirically validate whether feature selection enhances or limits the benefits of synthetic subsampling, seeking a technical balance between the computational complexity of the model and the diagnostic sensitivity required in real manufacturing environments. Based on the above, the overall objective of this research is to evaluate the performance of five machine learning algorithms under an experimental design consisting of six scenarios, obtained by combining three data management strategies (original distribution, SMOTE, and undersampling) with two levels of dimensionality (complete set of attributes and optimized feature selection) in order to identify the configuration with the best performance, capable of maximizing fault detection and minimizing false alarms in industrial environments that require high precision. This approach jointly integrates data imbalance management and relevant feature selection to ensure a robust, reliable model that is representative of the system's actual behavior. Based on the above, the following research question is formulated: Which of the six approaches evaluated, with data balancing and feature selection techniques, offers the most robust performance in fault detection, and what is the real impact of feature selection on model recall? Using the AI4I 2020 predictive maintenance dataset from the UCI Machine Learning Repository, this paper provides a methodical comparison assessment of five machine learning models for machine failure categorization. It shows how feature selection methods and data balance strategies (SMOTE and undersampling) impact predictive performance, showing that proper preprocessing can enhance problem detection while reducing false alarms. The study provides useful guidance for choosing the best model-data combinations, assisting in the creation of dependable and effective predictive maintenance systems in industrial settings where early fault detection and high accuracy are essential.

II. THEORETICAL REFERENCE

This section presents the conceptual foundations that underpin the proposed methodology, addressing the evolution of industrial maintenance, the principles of machine learning applied to classification, and the strategies used for processing unbalanced data, as well as the selection of relevant characteristics

II.1 FAILURE PREDICTION THEORY (PREDICTIVE MAINTENANCE)

Predictive Maintenance represents one of the main technological transformations in industrial maintenance within Industry 4.0. Unlike corrective and preventive maintenance, PdM is based on continuous monitoring of the condition of the machine with the aim of estimating its Remaining Useful Life and predicting failures before they occur [1]. In this context, technological convergence has evolved into what Bitam et al. (2025) define as AIoT (Artificial Intelligence of Things), enabling critical decentralization in fault detection [9]. From a theoretical point of view, it is assumed that mechanical failures do not occur suddenly or completely randomly, but are the result of progressive degradation processes. These processes are often reflected in measurable patterns within physical variables such as temperature, vibration, or torque, which can be detected prior to functional failure.

A recent systematic review by Guidotti et al. (2025) points out that the prevailing trend in the manufacturing industry is the adoption of data-driven models, due to their ability to capture nonlinear relationships that are often ignored by traditional analytical models [10]. Furthermore, Hector and Panjanathan (2024) emphasize that the effectiveness of these models depends not only on the algorithm, but also on the data planning and preparation process [11]. Finally, it is worth noting that studies such as that by Florian et al. (2021) show that the effective implementation of PdM can significantly reduce operating costs by eliminating unnecessary maintenance, transforming this activity from a cost center to a strategic element of industrial competitiveness [1].

II.2 MACHINE LEARNING THEORY

In Machine Learning, algorithms are mainly classified according to how they learn from data and the type of information available. There are six types, which are listed below: Supervised Learning, Unsupervised Learning, Semi-Supervised Learning, Reinforcement Learning, Deep Learning, Rule-Based Learning, and Hybrid Learning [12]. In this research, supervised learning algorithms are employed due to data availability, problem type (prediction), complexity of the environment, and computational resources. Among the most commonly used are Random Forest (RF), Decision Tree, Artificial Neural Networks (ANN), K-Nearest Neighbors (KNN), and Naive Bayes (NB). Machine learning focused on industrial fault diagnosis aims to build a mapping function capable of predicting a binary class target variable (fault and no fault) from a set of characteristics extracted from the system's sensors. The specialized literature emphasizes that the performance of the model depends largely on the correct choice of algorithm, which must consider the nature, dimensionality, and quality of the available data.

Research on complex datasets suggests that industrial environments require adaptive frameworks due to their irregular data, as validated by Autran et al. (2024) when working with synthetic and real maintenance data [13]. The Random Forest (RF) model is an ensemble method based on the Bagging (Bootstrap Aggregating) technique, which combines multiple decision trees to improve generalization ability. Recent comparative studies in industrial electrical systems have shown that RF outperforms models such as support vector machine and K-Nearest Neighbors in terms of accuracy and robustness, particularly in the presence of noise and diverse data [14]. Consistently, Yang and Wang (2025) confirm that RF is particularly effective in industrial environments characterized by high dimensionality and nonlinear relationships between variables [3]. Kale (2025) mentions that, to maximize accuracy in these models, it is important to optimize the hyperparameters of tree-based algorithms to capture representative patterns [15]. On the other hand, models such as Naive Bayes and Decision Trees (DT) continue to be relevant due to their simplicity and interpretability.

Naive Bayes, although based on the assumption of conditional independence between variables, is often used as a reference model, while DT facilitates the analysis of the decision-making process. However, Bezerra et al. (2024) warn that the performance of these algorithms depends critically on the quality of the input features, pointing out that indiscriminate elimination of variables during feature selection processes can degrade predictive power in complex industrial systems [7]. The performance evaluation of machine learning algorithms is based on indicators resulting from the confusion matrix. The common metrics are Accuracy, Recall, Precision, Specificity, and F1-Score, which are used to measure the effectiveness of the model by comparing or their predictions with actual data and verifying whether it is correctly classifying faults [10].

II.3 DATA BALANCING TECHNIQUES

One of the most significant theoretical challenges in failure classification is class imbalance, a phenomenon that gives rise to the called accuracy paradox, where a model can achieve high accuracy values without adequately detecting fault events due to their low statistical frequency. Among the most commonly used strategies is undersampling, which consists of randomly reducing the number of samples from the majority class. Although this approach reduces model bias, it carries the risk of losing relevant information. In contrast, the SMOTE (Synthetic Minority Over-sampling Technique) has established itself as one of the most widely used methodologies for addressing this problem. According to Yadav et al. (2025), from a mathematical point of view, SMOTE does not simply duplicate existing instances, but generates new synthetic samples by linear interpolation between a point in the minority class and its closest neighbors in the feature space [16]. However, Matharaarachchi et al. (2024) warn that oversampling loses effectiveness when there are outliers within the minority class. Their study proposes advanced variants that weight the proximity of the data to reduce the impact of these outliers and ensure that the synthetic samples generated are of high quality [17]. Recent research, such as that by Sridhar and Sanagavarapu (2021), shows that applying SMOTE to datasets from industrial sensors significantly improves model recall, allowing faults to be identified that would otherwise be misclassified as normal operating conditions [6].

III. MATERIALS AND METHODS

III.1 MACHINE LEARNING FRAMEWORK

The general methodology proposed for fault prediction is illustrated in Figure 1. This workflow integrates five machine learning models together with a cross-validation scheme to ensure the consistency of predictions. All experimental procedures were implemented using WEKA (3.8.6) [18].

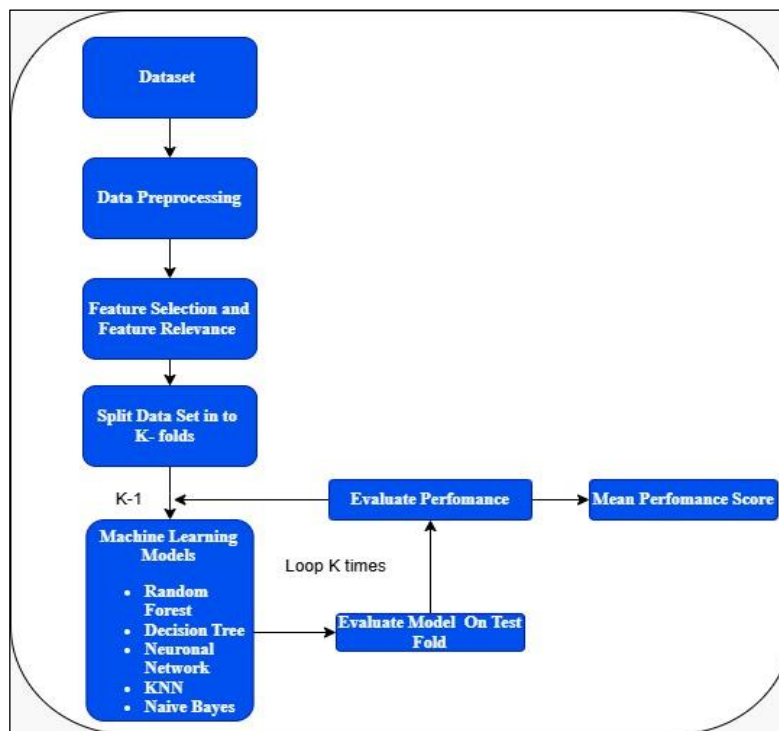


Figure 1: Machine learning framework for failure prediction.

Source: Authors, (2026).

III.1.1 Dataset Description

The study uses the AI4I 2020 Predictive Maintenance Dataset [5]. This synthetic dataset is modeled after an existing milling machine and consists of 10 000 data points from a stored as rows with 14 features in columns.

1. UID: unique identifier ranging from 1 to 10000.
2. Product ID: consisting of a letter L, M, or H for low (50% of all products), medium (30%) and high (20%) as product quality variants and a variant-specific serial number.
3. Type: just the product type L, M or H from column 2

4. Air temperature [K]: generated using a random walk process later normalized to a standard deviation of 2 K around 300 K
5. Process temperature [K]: generated using a random walk process normalized to a standard deviation of 1 K, added to the air temperature plus 10 K.
6. Rotational speed [rpm]: calculated from a power of 2860 W, overlaid with a normally distributed noise
7. Torque [Nm]: torque values are normally distributed around 40 Nm with a SD = 10 Nm and no negative values.
8. Tool wear [min]: The quality variants H/M/L add 5/3/2 minutes of tool wear to the used tool in the process.
9. Machine failure: label that indicates, whether the machine has failed in this particular datapoint for any of the following failure modes are true.

The machine failure consists of five independent failure modes:

10. Tool wear failure (TWF): the tool will be replaced or fail at a randomly selected tool wear time between 200 - 240 mins (120 times in our dataset). At this point in time, the tool is replaced 69 times, and fails 51 times (randomly assigned).
11. Heat dissipation failure (HDF): heat dissipation causes a process failure, if the difference between air- and process temperature is below 8.6 K and the tools rotational speed is below 1380 rpm. This is the case for 115 data points.
12. Power failure (PWF): the product of torque and rotational speed (in rad/s) equals the power required for the process. If this power is below 3500 W or above 9000 W, the process fails, which is the case 95 times in our dataset.
13. Overstrain failure (OSF): if the product of tool wear and torque exceeds 11,000 minNm for the L product variant (12,000 M, 13,000 H), the process fails due to overstrain. This is true for 98 datapoints.
14. Random failures (RNF): each process has a chance of 0,1 % to fail regardless of its process parameters. This is the case for only 5 datapoints, less than could be expected for 10,000 datapoints in our dataset.

If at least one of the above failure modes is true, the process fails and the 'machine failure' label is set to 1. It is therefore not transparent to the machine learning method, which of the failure modes has caused the process to fail. The original data distribution presents a class imbalance, with 9,661 instances of normal operation (96.6%) and only 339 instances of failure (3.4%). This disproportion represents a critical challenge for learning models.

III.2 DATA PREPROCESSING

Preprocessing was aimed at improving data quality and defining training strategies. The UDI and Product ID variables were eliminated because they lacked predictive value. Likewise, specific failure descriptors (TWF, HDF, PWF, OSF, RNF) were excluded, given that the study focuses on the binary detection of the machine failure variable. To address class imbalance, three experimental approaches are used:

- Approach 1 (SMOTE): The Synthetic Minority Over-sampling Technique (SMOTE) is applied to artificially balance the classes. In this scenario, synthetic data was generated until a total of 9,661 samples were reached for both classes. Furthermore, following the proposal by Matharaarachchi et al. [17], this technique improves model learning when minority classes exist, as is the case with the failure category in this study.
- Approach 2 (Undersampling): Random undersampling of the majority class is performed, resulting in a balanced set with 339 failure instances and 339 non-failure instances.
- Approach 3 (Original data set): Models are evaluated using the original data set with its original imbalance to establish a baseline.

III.3 FEATURE RELEVANCE AND SELECTION

Under these three approaches, a feature selection process is implemented by WEKA. The ReliefF evaluation algorithm (ReliefFAttributeEval) is used together with the Ranker search method. The Relief family of algorithms are highly successful attribute estimators due to their ability to detect conditional dependencies between variables, the algorithm evaluates the quality of each attribute by repeatedly sampling an instance and comparing it with its closest neighbors of the same class and of different classes. It assigns greater weight to characteristics that maintain similar values for instances of the same class and show different values for instances of different classes [19]. Establishing a cut-off criterion to select only the four most relevant characteristics for model training.

III.4 MACHINE LEARNING MODELS

For the training stage, a comparative approach is implemented to evaluate the predictive power of different machine learning models. Based on a review of recent literature by Aydın and Evrentuğ [20] and Arun [21], five representative algorithms (Random Forest, Decision Tree, ANN, KNN, and Naive Bayes) were selected that have proven effective in fault detection, as identified by Cioch [8]. The general operation of each implemented algorithm is described below: Random Forest (RF): This is an ensemble learning method based on the Bagging technique. It works by training multiple decision trees independently using random subsets of the data. For the final classification, the algorithm aggregates the results of all the individual trees and determines the winning class by majority vote, which reduces variance and improves robustness against noise compared to a single estimator [3]. Decision Tree (DT): This is an algorithm based on logical rules that models decisions in a hierarchical structure similar to a flowchart. The model recursively divides the data space into smaller regions, evaluating at each node the feature that offers the best separation of classes. Its main advantage, highlighted by Cioch et al. (2025), lies in the direct interpretability of the generated “If-Then” rules, although it can be prone to overfitting if its depth is not limited [8]. Artificial Neural Networks (ANN): A Multilayer Perceptron (MLP) is implemented, inspired by the biological structure of the brain. This algorithm processes information through layers of interconnected nodes (neurons): an input layer, hidden layers with nonlinear activation functions, and an output layer.

Yang and Iqbal (2025) explain that learning occurs through backpropagation, adjusting the weights of the connections to minimize error, which makes them powerful for capturing complex nonlinear patterns in computationally expensive environments [22]. K-Nearest Neighbors (KNN): This is an instance-based learning algorithm that does not construct an explicit internal model (Lazy Learning). To classify a new sample, it calculates its Euclidean distance from the points stored in the training set and identifies the ‘K’ nearest neighbors. Class assignment is performed by simple majority, which is an effective method when failures tend to cluster in specific regions of the vector space [10]. Naive Bayes (NB): This is a probabilistic classifier based on Bayes' theorem. Its operation assumes “naive independence” between predictor characteristics given the value of the class. Although Cioch et al. (2025) warn that this independence is rarely met in correlated sensor data, the model is widely used as a baseline due to its high computational efficiency and low memory consumption [8].

III.5 EVALUATION PERFORMANCE

The K-Fold Cross-Validation method is implemented for performance evaluation. This technique consists of randomly partitioning the total data set into K-Fold of equal or similar size, ensuring that the number of instances belonging to each class the same proportion in each partition, according to the methodology described by Lumumba [23]. As a result of this process, the confusion matrix defined by Cioch et al. [8] is obtained. Since this is a binary classification problem, the matrix is structured in a 2x2 dimension and its components are described as follows: TP (True Positives), the number of instances of class 1 (Failure) correctly predicted as class 1; TN (True Negatives), the number of instances of class 0 (Normal) correctly predicted as class 0; FP (False Positives), instances of class 0 incorrectly predicted as class 1; and FN (False Negatives), instances of class 1 incorrectly predicted as class 0 (Failure not detected). Based on the information obtained from this matrix, the performance metrics for the positive class described below are calculated. Accuracy: evaluates the overall performance of the model by measuring the proportion of correct predictions relative to the total number of cases.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (2)$$

Recall: Measures the ability of the algorithm to correctly identify all instances of actual failure.

$$Recall = \frac{TP}{TP + FN} \quad (3)$$

Specificity: Indicates the ability of the system to correctly classify normal operating states.

$$Specificity = \frac{TN}{TN + FP} \quad (4)$$

Precision: Determines the reliability of positive predictions of total failures.

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

F1-Score: Represents the harmonic mean between Precision and Recall, providing a critical balanced metric.

$$F1 = \frac{2 \times P \times R}{P + R} \quad (6)$$

IV. RESULTS AND DISCUSSION

In this section, the results obtained after the application of two data balancing approaches Original Imbalanced, SMOTE, and Undersampling—are presented and analyzed. These approaches are comprehensively evaluated using both the preprocessed feature set and after a variable selection process, allowing for the measurement of their performance on five machine learning models. Performance is quantified through a metrics analysis that includes accuracy, recall, specificity, precision, and F1-score, in order to determine the most robust configuration for fault detection in the industrial environment under study; as it is shown in Figure 2.

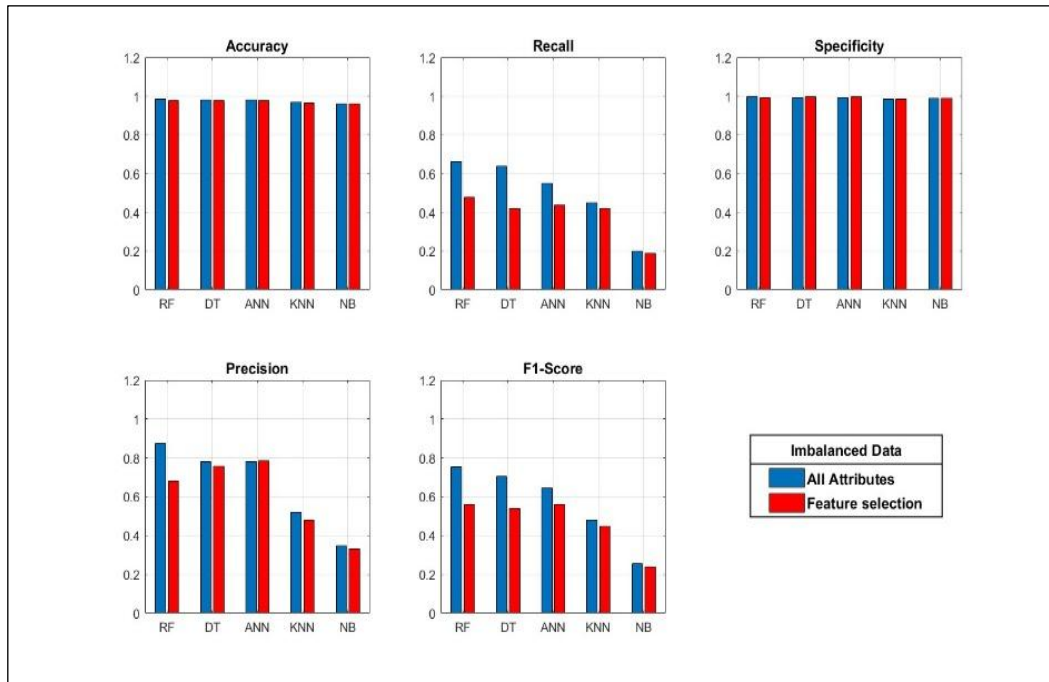


Figure 2: Performance comparison on the imbalanced dataset: all attributes vs feature selection.
Source: Authors, (2026).

When analyzing the results shown in Figure 2, one of the most critical challenges in predictive maintenance becomes evident: the bias introduced by imbalanced data. Although the models report high accuracy in both scenarios, this represents a significant issue; the imbalance causes the algorithms to become experts at classifying normal operation while losing effectiveness in identifying actual failures. This is clearly reflected in the low recall values, the metric that quantifies failure detection capability. Similarly, although the specificity shows good results, this is a product of the model's tendency to predict the majority class. When evaluating the individual performance of the algorithms, the Random Forest model stands as the most robust, specifically when operating with the complete feature set. Under this condition, the algorithm achieved the highest F1-Score in the comparison (0.7538), demonstrating a superior learning capacity to identify failure patterns compared to the other models. While algorithms such as Naive Bayes and KNN show a critical dependence on the majority class, Random Forest achieved the most balanced trade-off between event detection and alert reliability. On the other hand, it is essential to highlight the negative impact that feature selection had in this scenario. By reducing the dimensionality of the imbalanced dataset, a general decrease in the predictive capacity of all models is observed; for example, the Random Forest F1-Score drops drastically from 0.7538 to 0.5615. This behavior suggests that, given the scarcity of failure examples in the original dataset, the removing variables or features eliminates critical information that the algorithms need to correctly discriminate between operating states, confirming that model simplification can be counterproductive under imbalanced conditions. The performance metrics results for the synthetic data generated by the SMOTE technique are presented in Table 1.

Table 1: Performance metrics for the SMOTE-balanced dataset.

Model	Accuracy	Recall	Specificity	Precision	F1 Score
RF	0.9581	0.9663	0.95	0.9508	0.9585
DT	0.9435	0.9503	0.9367	0.9375	0.9439
ANN	0.9187	0.9424	0.895	0.8998	0.9206
KNN	0.9436	0.9585	0.9288	0.9308	0.9445
NB	0.8362	0.826	0.8464	0.8432	0.8345

Source: Authors, (2026).

Table 2: Performance metrics for the SMOTE-balanced dataset combined with Feature Selection.

Model	Accuracy	Recall	Specificity	Precision	F1 Score
RF	0.945	0.9619	0.9282	0.9305	0.9459
DT	0.9346	0.9445	0.9246	0.9261	0.9352
ANN	0.8823	0.9465	0.818	0.8387	0.8894
KNN	0.9114	0.9267	0.8961	0.8992	0.9127
NB	0.8275	0.7914	0.8636	0.853	0.821

Source: Authors, (2026).

Table 2 presents the performance metrics calculated for each model using a dataset balanced by SMOTE and feature selection. When analyzing the results in Tables 1 and 2, a positive change can be observed compared to the imbalanced scenario. The application of SMOTE to generate synthetic data has allowed the classes to be balanced, ensuring that the accuracy now truly reflects the model's performance. The most notable feature is the increase in Recall, where the Random Forest model reaches a maximum value of 0.9663 (Table 1). This indicates that the models have achieved reliable prediction of failures and have developed a real ability to identify critical predictive maintenance events without compromising specificity. In this approach, the Random Forest model once again stands out as the best performing model, especially when using the complete dataset. With an F1-Score of 0.9585 and the aforementioned Recall of 0.9663, Random Forest demonstrates high accuracy in detecting faults. This confirms that, by providing a balanced database, ensemble algorithms can extract much more complex and reliable patterns for decision-making. Regarding the impact of feature selection in this scenario, it can be observed that, although reducing variables simplifies the model structure, it leads to a slight decrease in performance. The clearest example is the Neural Network, where the F1-Score drops from 0.9206 to 0.8894; however, this deficiency does not compromise the overall effectiveness of the system. On the contrary, the results show that feature selection applied to a balanced dataset with SMOTE is effective, as the models manage to maintain a performance above 88% in fault detection. This confirms that the robustness of the synthetic database compensates for the elimination of technical variables, allowing for more efficient models without significantly sacrificing predictive power. Next, Tables 3 and 4 are presented, reflecting the dataset balanced to minority class size (undersampling) and the dataset balanced to minority class size (undersampling feature selection).

Table 3: Dataset balanced to minority class size (undersampling).

Model	Accuracy	Recall	Specificity	Precision	F1 Score
RF	0.9436	0.9585	0.9288	0.9308	0.9445
DT	0.9277	0.9381	0.9174	0.9191	0.9285
ANN	0.8997	0.8791	0.9204	0.9169	0.8976
KNN	0.9174	0.8997	0.9351	0.9327	0.9159
NB	0.8451	0.8407	0.8496	0.8482	0.8444

Source: Authors, (2026).

Table 4: Dataset balanced to minority class size (undersampling Feature selection).

Model	Accuracy	Recall	Specificity	Precision	F1 Score
RF	0.95	0.9528	0.9469	0.9472	0.95
DT	0.9381	0.9351	0.941	0.9407	0.9379
ANN	0.8938	0.8791	0.9086	0.9058	0.8922
KNN	0.9056	0.885	0.9263	0.9231	0.9036
NB	0.8245	0.7935	0.8555	0.8459	0.8189

Source: Authors, (2026).

Looking at the results in Tables 3 and 4, we can see an improvement in predictive power compared to the unbalanced dataset. By reducing the minority class, the models achieve good performance and a balance between accuracy and recall. In this approach, the Random Forest model stands out as the best-performing algorithm. Using the complete dataset, Random Forest achieves an F1-Score of 0.9445 and leads detection with a Recall of 0.9585. It can be noted that, although the dataset is smaller than in the case of SMOTE, the Random Forest algorithm maintains stability in its performance, consistently outperforming Neural Networks and models such as Naive Bayes in all metrics. Regarding the impact of feature selection seen in Table 4, a distinctive phenomenon occurs: unlike previous approaches, variable reduction helps improve or maintain performance in leading models. For example, in Random Forest, the F1-Score increases from 0.9445 to 0.9500 after applying Feature Selection. However, a slight deficiency is observed in models such as RNA, where performance drops from 0.8976 to 0.8922. The overall result shows that feature selection in a balanced dataset by subsampling is extremely effective, allowing for lighter and more accurate models.

V. CONCLUSIONS

This paper evaluates the performance of five machine learning algorithms for fault prediction in the AI4I 2020 industrial dataset, which allows us to understand the impact of each technique on critical metrics in industrial environments. To this end, data balancing and feature selection strategies are compared, leading to the following final considerations:

1. The optimal configuration: The Random Forest model, trained with the dataset balanced using the SMOTE technique and using all attributes after preprocessing, proves to be the most optimal solution, achieving a Recall of 0.9663 and an F1-Score of 0.9585. This result makes it the optimal method for fault prediction.
2. Impact of Preprocessing: It is demonstrated that treating class imbalance is a mandatory step. Using the original dataset, despite offering superior accuracy, results in models with significantly low sensitivities of less than 0.67. Among the balancing techniques, SMOTE consistently outperforms undersampling, proving that generating synthetic failure data is preferable to removing information from the majority class.
3. This research reveals that feature selection negatively affects predictive performance, reducing the fault detection capability (Recall) in the best-performing model. This suggests that the operational variables in the dataset are significantly relevant to the models and are lost when those attributes are removed, indicating that for this specific problem, it is better to retain the attributes.

Finally, the integration of data balancing and feature selection improves overall performance because the results show that treating class imbalance combined with feature selection techniques allows for the construction of more robust and stable models. The joint integration of these strategies contributes to improving the predictive capacity of the system compared to their isolated application.

VI. AUTHOR'S CONTRIBUTION

Conceptualization: Juan Carlos Gonzalez Islas, Jesus Vargas-Ortiz, Aldo Márquez-Grajales, Ernesto Bolaños-Rodríguez, Asdrúbal López-Chau, Rosa María Ortega-Mendoza.

Methodology: Juan Carlos Gonzalez Islas, Jesus Vargas-Ortiz, Aldo Márquez-Grajales, Ernesto Bolaños-Rodríguez, Asdrúbal López-Chau, Rosa María Ortega-Mendoza.

Investigation: Juan Carlos Gonzalez Islas, Jesus Vargas-Ortiz, Aldo Márquez-Grajales, Ernesto Bolaños-Rodríguez, Asdrúbal López-Chau, Rosa María Ortega-Mendoza.

Discussion of results: Juan Carlos Gonzalez Islas, Jesus Vargas-Ortiz, Aldo Márquez-Grajales, Ernesto Bolaños-Rodríguez, Asdrúbal López-Chau, Rosa María Ortega-Mendoza.

Writing – Original Draft: Juan Carlos Gonzalez Islas, Jesus Vargas-Ortiz, Aldo Márquez-Grajales, Ernesto Bolaños-Rodríguez, Asdrúbal López-Chau, Rosa María Ortega-Mendoza.

Writing – Review and Editing: Juan Carlos Gonzalez Islas, Jesus Vargas-Ortiz, Aldo Márquez-Grajales, Ernesto Bolaños-Rodríguez, Asdrúbal López-Chau, Rosa María Ortega-Mendoza.

Resources: Juan Carlos Gonzalez Islas, Jesus Vargas-Ortiz, Aldo Márquez-Grajales, Ernesto Bolaños-Rodríguez, Asdrúbal López-Chau, Rosa María Ortega-Mendoza.

Supervision: Juan Carlos Gonzalez Islas, Jesus Vargas-Ortiz, Aldo Márquez-Grajales, Ernesto Bolaños-Rodríguez, Asdrúbal López-Chau, Rosa María Ortega-Mendoza.

Approval of the final text: Juan Carlos Gonzalez Islas, Jesus Vargas-Ortiz, Aldo Márquez-Grajales, Ernesto Bolaños-Rodríguez, Asdrúbal López-Chau, Rosa María Ortega-Mendoza.

VII. DATA AVAILABILITY

The data presented in this study are available upon request from the corresponding author.

VIII. REFERENCES

- [1] E. Florian, F. Sgarbossa, e I. Zennaro, "Machine Learning-Based Predictive Maintenance: A Cost-Oriented Model for Implementation," *International Journal of Production Economics*, vol. 236, p. 108114, 2021. doi: 10.1016/j.ijpe.2021.108114.
- [2] C. Tsallis, P. Papageorgas, D. Piromalis, y R. A. Munteanu, "Application-Wise Review of Machine Learning-Based Predictive Maintenance: Trends, Challenges, and Future Directions," *Applied Sciences*, vol. 15, no. 9, p. 4898, 2025. doi: 10.3390/app15094898.
- [3] Y. Yang y H. Wang, "Random Forest-based machine failure prediction: A performance comparison," *Applied Sciences*, vol. 15, no. 16, p. 8841, 2025. doi: 10.3390/app15168841.
- [4] K. I. Romero Magdaleno, L. C. Méndez González, I. J. González Hernández, y A. E. Quezada Carreón, "Implementación de modelos de aprendizaje automático para el mantenimiento predictivo y clasificación de fallas de maquinaria industrial," *Ingenio y Conciencia Boletín Científico de la Escuela Superior Ciudad Sahagún*, vol. 12, no. Especial, pp. 56–67, 2025..
- [5] S. Matzka, "Explainable Artificial Intelligence for Predictive Maintenance Applications," in *Proceedings of the 2020 Third International Conference on Artificial Intelligence for Industries (AI4I)*, Irvine, CA, USA, 2020, pp. 69–74, doi: 10.1109/AI4I49448.2020.00023.
- [6] S. Sridhar y S. Sanagavarapu, "Handling data imbalance in predictive maintenance for machines using SMOTE-based oversampling," en *2021 13th International Conference on Computational Intelligence and Communication Networks (CICN)*, 2021, pp. 44-49. doi: 10.1109/CICN51697.2021.9574668.
- [7] F. E. Bezerra et al., "Impacts of feature selection on predicting machine failures by machine learning algorithms," *Applied Sciences*, vol. 14, no. 8, p. 3337, 2024. doi: 10.3390/app14083337.
- [8] M. Cioch, M. Kulisz, and A. Gola, "Comparison of Machine Learning Methods in Predictive Maintenance of Machines," *Advances in Science and Technology Research Journal*, vol. 19, no. 11, pp. 33-44, 2025. doi: 10.12913/22998624/208284.
- [9] T. Bitam et al., "Artificial Intelligence of Things for Next-Generation Predictive Maintenance," *Sensors*, vol. 25, no. 24, p. 7636, 2025. doi: 10.3390/s25247636.
- [10] D. Guidotti, L. Pandolfo, y L. Pulina, "A systematic literature review of supervised machine learning techniques for predictive maintenance in Industry 4.0," *IEEE Access*, vol. 13, pp. 102479-102504, 2025. doi: 10.1109/ACCESS.2025.3578686.
- [11] I. Hector y R. Panjanathan, "Predictive maintenance in Industry 4.0: a survey of planning models and machine learning techniques," *PeerJ Computer Science*, vol. 10, p. e2016, 2024. doi: 10.7717/peerj-cs.2016.
- [12] Q. Wan, "Machine Learning Model Construction and Practice for Personalized Training Programs in Physical Education and Sport Teaching," *Applied Mathematics and Nonlinear Sciences*, vol. 10, no. 1, 2025. doi: 10.2478/amns-2025-0374.
- [13] J.-V. Autran, V. Kuhn, J.-P. Diguët, M. Dubois, y C. Buche, "AI4I-PMDI: Predictive maintenance datasets with complex industrial settings' irregularities," *Procedia Computer Science*, vol. 246, pp. 1201-1209, 2024. doi: 10.1016/j.procs.2024.09.546.
- [14] S. Singh et al., "A comparative analysis of machine learning algorithms for predictive maintenance in electrical systems," en *2024 4th International Conference on Innovative Practices in Technology and Management (ICIPTM)*, 2024, pp. 1-6. doi: 10.1109/ICIPTM59628.2024.10563826.
- [15] S. S. Kale, "Predictive Maintenance of Equipment using Machine Learning Algorithms," *Tesis Doctoral*, National College of Ireland, Dublin, 2025.

- [16] S. Yadav et al., "CSRBoost: Clustered Sampling With Resampling Boosting for Imbalanced Dataset Pattern Classification," *IEEE Access*, vol. 13, pp. 175078-175093, 2025. doi: 10.1109/ACCESS.2025.3616207.
- [17] S. Matharaarachchi, M. Domaratzki, y S. Muthukumarana, "Enhancing SMOTE for imbalanced data with abnormal minority instances," *Machine Learning with Applications*, vol. 18, p. 100597, 2024. doi: 10.1016/j.mlwa.2024.100597.
- [18] I. H. Witten, E. Frank, M. A. Hall, and C. J. Pal, *Data Mining: Practical Machine Learning Tools and Techniques*, 4th ed. Burlington, MA, USA: Morgan Kaufmann, 2016. doi: 10.1016/C2009-0-19715-5.
- [19] M. Robnik-Šikonja and I. Kononenko, "Theoretical and Empirical Analysis of ReliefF and RReliefF," *Machine Learning*, vol. 53, no. 1, pp. 23–69, 2003. doi: 10.1023/A:1025667309714.
- [20] C. Aydın y B. Evrentuğ, "Evaluation of predictive maintenance efficiency with the comparison of machine learning models in machining production process in brake industry," *PeerJ Computer Science*, vol. 11, p. e2999, 2025. doi: 10.7717/peerj-cs.2999.
- [21] K. Arun, "Evaluating Machine Learning Algorithms for Predictive Maintenance in Equipment," *Global J. Eng. Innov. Interdiscip. Res.*, vol. 5, 2025. doi: 10.33425/3066-1226.1059.
- [22] Y. Yang and M. Z. Iqbal, "Cost-Optimised Machine Learning Model Comparison for Predictive Maintenance," *Electronics*, vol. 14, no. 12, p. 2497, 2025. doi: 10.3390/electronics14122497.
- [23] V. Lumumba et al., "Comparative analysis of cross-validation techniques: LOOCV, k-fold cross-validation, and repeated k-fold cross-validation in machine learning models," *Am. J. Theor. Appl. Stat.*, vol. 13, pp. 127–137, 2024. doi: 10.11648/j.ajtas.20241305.13.