Predictive Maintenance in Manufacturing: A Comparative Analysis of K-Means Clustering and Random Forest

Nicholas Franck
Department of Computer Science
University of South Carolina – Upstate
Spartanburg, South Carolina
NFranck@email.uscupstate.edu

ABSTRACT

Maintenance is an important and essential process in manufacturing. With constant moving parts and wear and tear on machinery, maintenance becomes necessary. As one of manufacturing's largest operational cost, the need for cost reduction is a requirement for increased profits. A good way to reduce the cost is a process called predictive maintenance. This is done by trying to predict the lifespan of production equipment and repairing or replacing the equipment when the production line is not in operation. Predictive maintenance is a topic that allows for maintenance to reduce its operational cost, to predict when machinery will fail. By doing this, two outcomes are produced: the machinery gets the most out of its lifespan and machines are producing longer, decreasing downtime. Machine learning is a great way to help predict the best time to replace the machinery. Two machine learning algorithms that can be used are K-means cluster and Random Forest Classification. This study will use these two machine learning models in order to determine their efficacy and use in practical predictive maintenance tasks in a manufacturing environment. These models are selected in order to determine if either a supervised learning model, Random Forest, or an unsupervised learning model, K-means, performs better for the given task.

Following the completion of the project, it was observed that the Random Forest Classification algorithm exhibited superior performance in predicting machine failure compared to the K-means cluster algorithm. The findings suggest that the supervised learning approach of Random Forest outperformed the unsupervised learning model, highlighting its efficacy in practical predictive maintenance tasks within a manufacturing environment. Emphasizing the potential of Random Forest Classification as a valuable tool for enhancing maintenance strategies and reducing operational costs in manufacturing processes.

Keywords

Machine Learning, ML, K-Means Cluster, K-Means, Random Forest, Random Forest Classification, Maintenance, Predictive Maintenance, Manufacturing, Predictive Maintenance in Manufacturing

1. INTRODUCTION

Maintenance is a fundamental requirement in manufacturing, as machinery is a very important part of production processes. However, the cost of maintenance can be a significant operational burden, impacting the core of manufacturing industries. To address this challenge and boost profitability, the concept of predictive maintenance has become an important field of study.

Predictive maintenance is a proactive strategy that seeks to forecast machinery failures. By using data-driven analytics, it enables timely maintenance, optimizing the cost control and operational efficiency. The benefits are machinery operating at its maximum potential lifespan, and production lines experiencing minimal downtime.

In this research project, I delve into predictive maintenance, specifically focusing on the role of machine learning algorithms in predicting manufacturing machine failures. Two prominent machine learning techniques: K-means clustering, and Random Forest Classification are the focus of this research. Through an analysis of these approaches, I aim to provide insights into their use for predictive maintenance applications in manufacturing.

The outcome of this research is significant for manufacturing industries seeking to increase profitability through predictive maintenance. The ability to select the most effective model can translate into substantial cost savings, improved machinery reliability, and uninterrupted production schedules.

The significance of this research extends to manufacturing industries worldwide. In an era marked by an increasing emphasis on operational efficiency, cost containment, and enhanced machinery reliability, the findings of this research are poised to make large contributions. By offering a great understanding of machine learning models such as K-means clustering and Random Forest Classification, this research equips manufacturers with the knowledge to make data-driven decisions. Selecting the most suitable predictive maintenance model, as shown by this study, holds the potential to yield great benefits, including substantial cost savings, improved machinery reliability, and seamless, uninterrupted production schedules.

2. LITERARY REVIEW

Predictive maintenance has become a critical part of modern manufacturing, offering the potential to increase efficiency, reduce costs, and minimize downtime. In this literature review, I explore other research, and findings related to predictive maintenance, with a focus on the role of machine learning algorithms.

Breiman's work on Random Forests has left a mark on the field of machine learning. One key finding from this work is the robustness of Random Forests in handling complex datasets and their ability to reduce overfitting, resulting in improved prediction

accuracy. These attributes make Random Forests Classification a good choice for predictive maintenance applications [1].

Rodriguez conducted an analysis of data from wind turbines, utilizing K-Means clustering to detect patterns indicative of maintenance needs. Their findings show the efficacy of K-Means clustering in identifying distinct machinery behavior clusters. This allows for proactive maintenance interventions based on the observed patterns, potentially reducing maintenance costs and downtime [7].

Ouadah researched algorithm selection for predictive maintenance. Their study highlights the importance of choosing the most appropriate supervised machine learning algorithm for a specific application. One notable finding is that the selection process significantly impacts predictive maintenance model performance, emphasizing the need for a good algorithm choice [6].

Li researched the effects of predictive maintenance on transportation networks, particularly in the rail industry. Their research findings revealed the potential of machine learning to improve rail network velocity. By proactively identifying maintenance needs, this approach could lead to smoother operations, reduced disruptions, and enhanced overall efficiency [2].

In summary, these research findings show the growing interest of predictive maintenance and its role in manufacturing. Machine learning techniques, including Random Forests and K-Means clustering, offer valuable tools for optimizing maintenance strategies. The takeaway of these research is that a well-informed selection of machine learning algorithms can significantly impact predictive maintenance model performance, leading to improved operational efficiency and cost savings across various industries. As industries continue to embrace data-driven approaches, predictive maintenance remains a viable solution to reducing manufacturing costs.

3. METHODOLOGY

This section provides a detailed insight into the methodology, encompassing the application of K-Means Clustering and Random Forest Classification, with the goal of improving predictive maintenance in manufacturing.

3.1 K-Means Clustering

K-Means Clustering could become an important technique in predictive maintenance within the manufacturing industry, as it assists in identifying patterns in machinery data, enabling timely maintenance on machinery.

The initial phase involves data preprocessing, the collection of relevant data, cleansing, and normalization. This step ensures high-quality input data, a fundamental requirement for effective clustering.

A significant aspect of implementing K-Means Clustering is determining the optimal number of clusters, denoted as 'k.' This is achieved through the elbow method, a technique that entails plotting the sum of squared distances against different 'k' values. The 'elbow' point on the graph, where the rate of decrease sharply changes, signifies the optimal 'k' value [4].

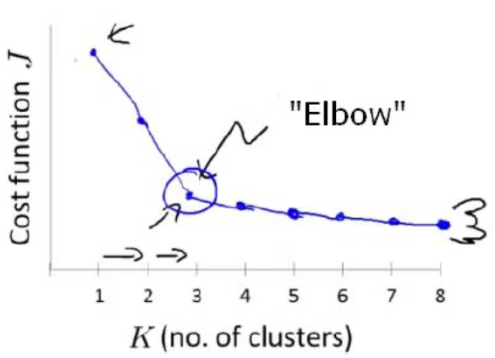


Figure 1: Elbow Curve [4].

Additionally, the K-Means algorithm is implemented with the determined 'k' value, effectively segmenting the data into clusters. These clusters represent behavior patterns.

A key component in the K-Means Clustering process is cluster analysis. This phase entails an in-depth examination of each cluster to reveal distinctive behavior patterns within data.

Furthermore, to assess the quality of the clustering results, the silhouette score is applied. It evaluates how close each data point is to its own cluster compared to other clusters. Higher silhouette scores indicate well-separated clusters, providing valuable information for predictive maintenance decision-making [4].

K-means clustering has several advantages that make it a popular choice for data analysis. It's known for its simplicity, efficiency, and ease of implementation, making it suitable for large datasets. K-means is highly interpretable, scalable, and adaptable to different data types, including numerical and categorical variables. It offers immediate results and works well for compact, evenly sized, and evenly distributed clusters.

However, K-means clustering also comes with limitations. It's sensitive to the initial placement of centroids and may produce divergent results with different initializations. The algorithm requires the number of clusters (K) to be predefined, which can be challenging. K-means assumes equally sized clusters, which may not always hold in real-world data. It's also sensitive to outliers and less suitable for irregularly shaped clusters.

3.2 Random Forest Classification

Random Forest Classification is another possible tool that could be used in predictive maintenance for the manufacturing industry. This methodology involves a series of steps to effectively predict machinery failures.

Data preprocessing is the foundational step, encompassing data collection, cleansing, and normalization. This ensures high-quality input data essential for accurate predictions. Feature selection is equally important, identifying the most relevant attributes for classification [1].

The Random Forest model, composed of multiple decision trees, is constructed using the training data. Each tree operates independently on different data subsets and features, reducing overfitting.

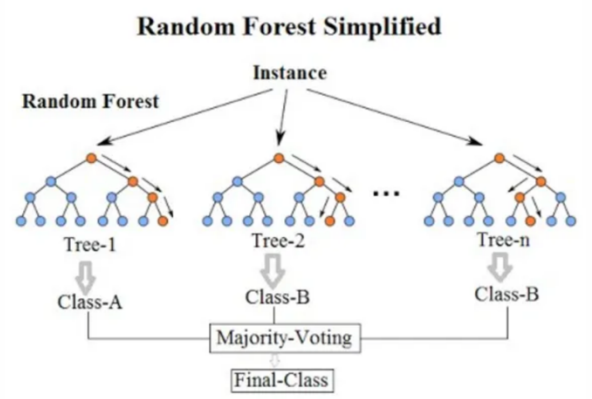


Figure 2: Multiple Decision Trees [3].

Random Forest allows customization of decision tree parameters, including depth and feature selection. Model performance is evaluated using metrics like accuracy, precision, recall, F1 score, and ROC curves, gauging its ability to distinguish machinery failures from non-failures.

Random Forest ranks features by importance, shedding light on key factors contributing to failure. Cross-validation techniques and hyperparameter tuning are applied to assure the model's robustness and tune performance [1].

Random Forest excels in resistance to overfitting, handling high-dimensional data, and feature importance ranking. However, hyperparameter tuning is necessary for optimal performance, and resources must be managed, especially with a large number of trees.

4. IMPLEMENTATION

4.1 Technology

The implementation of predictive maintenance models was carried out within a Python environment, with PyCharm being the IDE of choice. Python was chosen for its compatibility with key libraries and packages crucial for data analysis and machine learning. Pandas played a role in managing the dataset, allowing for the import of CSV files, data frame creation, and data manipulation during the preprocessing phase. Scikit-Learn, previously known as Sklearn, was used in scaling the data, training the models, and assessing their performance, making it a useful tool for implementing machine learning algorithms. Matplotlib facilitated data visualization, enabling the creation of plots and graphs, including the elbow curve used in model evaluation.

4.2 Data

The dataset comprises of 10,000 data points, each uniquely identified by a UID ranging from 1 to 10,000. It includes product IDs denoting product quality variants (L, M, H) and product types (L, M, H). Other attributes encompass air temperature, process temperature, rotational speed, torque, tool wear duration, and a binary 'machine failure' label. This label indicates the occurrence of a machine failure, which may result from five distinct failure modes: tool wear failure (TWF), heat dissipation failure (HDF), power failure (PWF), overstrain failure (OSF), and random failures (RNF). The dataset corresponds to the publication by S. Matzka, titled "Explainable Artificial Intelligence for Predictive Maintenance Applications" [8].

5. EXPERIMENTAL SETUP

5.1 Data Preprocessing

5.1.1 Column Standardization

In the initial phase of data preprocessing, column standardization was performed to create uniformity in the dataset. This process includes converting all column headers into a consistent format, replacing spaces with underscores, and ensuring that each column adhered to a standard naming convention. By unifying the structure of column names, the dataset was optimized for later analyses, allowing for easier coding, and consistency throughout the research.

5.1.2 Irrelevant Columns

To streamline the dataset, a selection of columns were identified as irrelevant and were removed. These columns include the unique identifier and product types, which held limited relevance to the core objectives of the study. Additionally, all columns associated with the five independent failure modes were excluded, as the reason for failure is not the cause for this research. By eliminating these irrelevant columns, the dataset was refined, enhancing its feature set for predictive maintenance investigations.

5.1.3 'machine_failure' Column

The 'machine_failure' column constitutes the objective of this research – predicting the occurrence of machine failures. This column is marked '1' for failure and '0' for no failure, it serves as the target variable for predictive maintenance. To produce accurate predictions, this column is removed from the feature set and assigned as the outcome to be predicted.

5.1.4 Feature Columns

The feature columns, encompassing air temperature, process temperature, rotational speed, torque, and tool wear duration, represent the fundamental predictors instrumental in our predictive maintenance model. These attributes encapsulate critical process parameters and contextual data vital for the assessment and prediction of machine failures. The air temperature and process temperature provide insights into the thermal dynamics of the machining process, while rotational speed and torque offer key indicators of the mechanical aspects. Additionally, the tool wear duration serves as an essential variable, influencing the overall health of the machinery. These feature columns play an important role in shaping the predictive capabilities of our model, allowing to detect early warning signs of impending failures and undertake any needed maintenance actions.

5.1.5 Training Split

To evaluate the machine learning models, I divided the dataset into two distinct subsets: the training set and the testing set. The training set, encompassing 80% of the data, serving as the foundation for model training. During this phase, the models gain insights into historical data, allowing them to discern complex patterns and relationships. The remaining 20% of the dataset, designated as the testing set, remained unseen during model

training. This partitioning strategy is crucial to ascertain the models' robustness, preventing issues like overfitting or underfitting, and ensuring their reliability in practical manufacturing predictive maintenance scenarios.

5.2 Algorithm Implementation

After preprocessing the data, the data can now be sent through the algorithms for training and then further for testing.

5.2.1 K-Means Clustering

The application of K-Means clustering for predictive maintenance involves first separating the groups into the two types of prediction: the “Machine Failure” group and the “No Machine Failure” group. For both groups, the elbow method and silhouette scoring are employed to discern the optimal number of clusters for each group.

After the number of clusters is known, both groups can be clustered around centroids. Once they have been clustered, the testing data is input, and each point is tested to find the distance from each centroid. Whichever centroid has the closest distance is where the point belongs, and whichever centroid the group belongs to is the prediction for the testing point. This is applied across all the test data, compiling a list of predictions.

5.2.2 Random Forest Classification

This application of Random Forest Classification for predictive maintenance involves first applying hyperparameter tuning, which is not necessary, however it does improve precision. After this is done, the hyperparameters are used to fit the Random Forest model. Then the test data can be input through the trained model and predictions can be assessed.

5.3 Evaluation Metrics

After both models have been tested against the training data, they are evaluated on a set of metrics to determine the efficacy of its use in predictive maintenance.

5.3.1 K-Means Clustering

In the evaluation of K-Means Clustering's performance for predictive maintenance, two unique and one shared evaluation metrics are employed.

The Silhouette Score quantifies the cohesion and separation of data points within clusters, with a higher score indicating well-defined clusters. This metric provides insights into the effectiveness of K-Means in identifying machinery behavior patterns [4].

Additionally, the Elbow Method to determine the optimal number of clusters (k) by analyzing the sum of squared distances for different k values. The "elbow" point on the graph indicates the optimal k value, offering granularity insights into the clustering process [4].

Lastly, the prediction accuracy for the model will be used to determine overall how likely is it to predict machine failures correctly. This will be the main comparison metric with Random Forest, as it is the only shared metric between the two.

5.3.2 Random Forest Classification

In the evaluation of Random Forest Classification, a much wider set of metrics than K-means are needed to get a detailed analysis of its results and efficacy.

Accuracy, which measures the proportion of correctly classified instances, offers an overall performance assessment.

Precision assesses the model's accuracy in classifying true positive instances among all instances predicted as positive, making it crucial when false positives come at a high cost [1].

Recall evaluates the model's ability to identify true positive instances among all actual positive instances, particularly important when missing a true positive incurs a high cost [1].

The Receiver Operating Characteristic-Area Under the Curve (ROC-AUC) quantifies the model's ability to distinguish between failure and non-failure instances, offering insights into its performance [5].

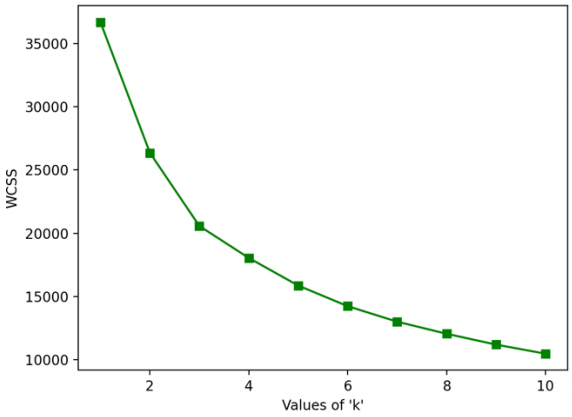
6. RESULTS AND ANALYSIS

This section unfolds the outcomes of the exploration into predictive maintenance using K-Means Clustering and Random Forest Classification. Diving into the performance metrics, dissecting how each model fares in foreseeing machine failure.

6.1 K-Means Clustering Results

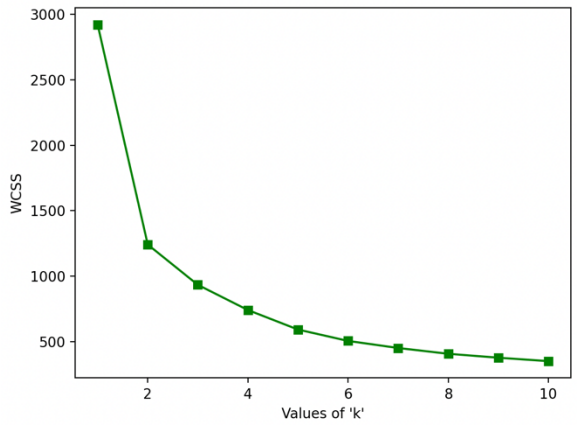
Since K-Means involves a first step of finding the optimal clusters the results include using the elbow method for each group, silhouette scoring each group, and finally the prediction statistics.

6.1.1 No Machine Failure Elbow Method



The elbow method for the “No Machine Failure” group shows no clear “elbow” which would support an optimal number of clusters for K-means. As the slope from 2 to 3 decreases abnormally from 3 to 4 it could be argued that there is an “elbow”, however it is slight and not enough to make a declaration of an optimal “k” clusters.

6.1.2 Machine Failure Elbow Method



The elbow method for "Machine Failure" group shows a clear "elbow" at 2 as the slope from 1 to 2 has a sharp change from 2 to 3. Making 2 the optimal "k" value from the above "elbow" method. However, what is not considered is the sample size that of the "No Machine Failure" group is much higher than the "Machine Failure" group, therefore 2 clusters will not be enough to offset this difference.

6.1.3 Cluster Silhouette Scores

Silhouette Scores		
# of Clusters	No Machine Failure	Machine Failure
2	0.265	0.655
3	0.264	0.344
4	0.234	0.325
5	0.228	0.338
6	0.225	0.354
7	0.225	0.338
8	0.227	0.311
9	0.222	0.309
10	0.222	0.292

The Silhouette Scores demonstrate how pure the clusters would form for each group based on the number of clusters being used. For the "No Machine Failure" group the purity score is very close for each amount with the highest score being for 2 clusters. For the "Machine Failure" group, with 2 clusters it has by far the highest purity score. However, this is not considering the difference in sample sized between this group and the "No Machine Failure" group therefore the number of clusters for each group need to be scaled.

6.1.4 K-Means Prediction Statistics

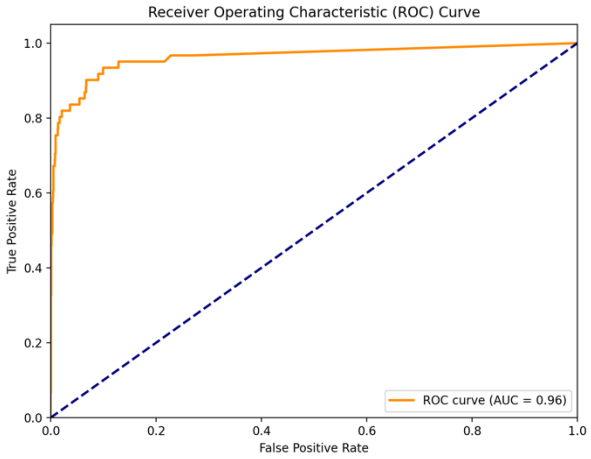
Predictions	Precision	Recall	Total Weighted Accuracy
No Machine Failure	97%	99%	95%
Machine Failure	32%	13%	

By scaling the cluster amounts the best-case prediction results are output. The algorithm is 97% precise when predicting "No Machine Failure" with 99% recall. However, it is only 32% precise with 13% recall when predicting "Machine Failure". Making the prediction's weighted accuracy to be 95%.

6.2 Random Forest Results

For Random Forest Classification the results only include the Receiver Operating Characteristic-Area Under the Curve and the prediction statistics.

6.2.1 Random Forest ROC-AUC



In the evaluation of Random Forest Classification, the Receiver Operating Characteristic-Area Under the Curve (ROC-AUC) demonstrates a notable performance, yielding a curve with an impressive AUC value of 0.96. This metric reflects the model's robust ability to distinguish between failure and non-failure instances.

6.2.2 Random Forest Prediction Statistics

Predictions	Precision	Recall	Total Weighted Accuracy
No Machine Failure	99%	100%	98%
Machine Failure	83%	57%	

The prediction shows that when using Random Forest Classification, with hyperparameter tuning, the total weighted accuracy is 98%. With the precision being 99% with 100% recall for the "No Machine Failure" prediction. And it is 83% precise with 57% recall for the "Machine Failure" prediction.

6.3 Analysis of Results

In order to have a clear understanding of what the data provided from the results mean it is imperative to analyze and compare the results.

6.3.1 K-Means Cluster Results Analysis

The analysis of the K-means results reveals a notable limitation in its performance, especially in the context of predicting machine failures. While the overall accuracy appears high, a closer examination uncovers a significant skew in the precision accuracy, predominantly driven by the "No Machine Failure" predictions with a substantially larger sample size. The primary objective of this project centers on the accuracy of predicting failures, and the findings indicate that K-means struggles in this specific aspect. The model tends to exhibit poor predictive capabilities for "Machine Failure", frequently generating false positives and indicating failure as shown by the poor recall results. This low recall was expected as the purity scores of the "No Machine Failure" were quite low causing incorrect predictions for "Machine Failure".

6.3.2 Random Forest Results Analysis

The analysis of the Random Forest results shows its great performance in terms of overall prediction accuracy, notably achieving a high level of precision in “Machine Failure” predictions. The model exhibits a strong ability to correctly identify instances of machine failures, contributing to its effectiveness as a predictive maintenance tool. However, a noteworthy aspect of the analysis reveals a limitation in the form of incorrect predictions of “Machine Failure” as shown by the recall. This means that it over predicts the amount of machine failures around 50% of the time. Even with such a mediocre recall from the “Machine Failure” predictions, the ROC curve had an impressive score of 96%, however this is again caused by the oversaturation of samples from the “No Machine Failure” predictions.

6.3.3 Results Comparison

While both models achieved high overall accuracy, Random Forest Classification outperforms K-Means clustering in precision for predicting machine failures. K-Means clustering excelled in accurately predicting instances of "No Machine Failure" but struggled with precision in the critical task of predicting actual machine failures. Random Forest Classification demonstrated a more balanced precision, making it a promising choice for industries where accurate identification of machine failures is paramount.

7. CONCLUSION

In conclusion, this research project delves into the subject of predictive maintenance in manufacturing, specifically implementing K-Means clustering and Random Forest Classification. The findings show the strengths and limitations of each model, providing important insights for the application of machine learning in predictive maintenance.

K-Means clustering, despite its simplicity and efficiency, reveals challenges when faced with imbalanced datasets. This limitation hinders its performance in predicting machine failures, as the model tends to favor the dominant class, thereby compromising its ability to accurately detect failures. The oversaturation of "No Machine Failure" samples significantly impacted the results, contributing to the model’s suboptimal performance in predicting actual machine failures. This highlights the necessity for more ‘supervised’ approaches, particularly in scenarios involving skewed data distributions.

On the contrary, Random Forest Classification emerges as a promising tool for predictive maintenance, boasting robust precision and accuracy. The model excels in correctly identifying machine failures, showcasing great performance and potential. However, it comes with a trade-off, as it tends to overpredict failures. This can be counterproductive since the primary objective of predictive maintenance is to minimize machinery costs by maximizing the lifespan of equipment and conducting maintenance just before it becomes necessary. Even with this deficit, Random Forest Classification still shows itself as an excellent tool for predictive maintenance.

In the larger context of manufacturing, the insights gained from this research hold significance for informed data driven decision making. The choice between K-Means clustering and Random Forest Classification should be tailored to the specific requirements of the manufacturer, considering factors such as dataset characteristics and the importance of accurate machine failure predictions.

Ultimately, this research contributes to the growing landscape of predictive maintenance knowledge, guiding manufacturing industries toward the selection of the most suitable models. As technology advances and more data becomes available, ongoing improvements of predictive maintenance strategies remains imperative. Continuous improvements in operational efficiency, cost reduction, and uninterrupted production schedules depend on the informed data driven decisions of advanced machine learning models in manufacturing.

8. REFERENCES

[1]. Breiman, L. Random Forests. Machine Learning 45, 5–32 (2001). https://doi.org/10.1023/A:1010933404324

[2]. Li H, Parikh D, He Q, et al. Improving rail network velocity: A machine learning approach to predictive maintenance. Transportation Research Part C: Emerging Technologies. 2014;45:17-26. doi:https://doi.org/10.1016/j.trc.2014.04.013

[3]. K, G. M. (2020, September 22). Machine learning basics: Random Forest classification. Medium. https://towardsdatascience.com/machine-learning-basics-random-forest-classification-499279bac51e

[4]. Kodinariya, T., Makwana, P. R. (2013, November 6). Review on determining of cluster in K-means clustering. International Journal of Advance Research in Computer Science and Management Studies. https://www.researchgate.net/profile/Trupti-Kodinariya/publication/313554124_Review_on_Determining_of_Cluster_in_K-means_Clustering/links/5789fda408ae59aa667931d2/Review-on-Determining-of-Cluster-in-K-means-Clustering.pdf

[5]. Narkhede, S. (2018, June 26). *Understanding AUC - ROC Curve*. 48hours. https://48hours.ai/files/AUC.pdf

[6]. Ouadah, A., Zemmouchi-Ghomari, L. & Salhi, N. Selecting an appropriate supervised machine learning algorithm for predictive maintenance. Int J Adv Manuf Technol 119, 4277–4301 (2022). https://doi.org/10.1007/s00170-021-08551-9

[7]. Rodriguez PC, Marti-Puig P, Caiafa CF, Serra-Serra M, Cusidó J, Solé-Casals J. Exploratory Analysis of SCADA Data from Wind Turbines Using the K-Means Clustering Algorithm for Predictive Maintenance Purposes. Machines. 2023; 11(2):270. https://doi.org/10.3390/machines11020270

[8]. S. Matzka, "Explainable Artificial Intelligence for Predictive Maintenance Applications," 2020 Third International Conference on Artificial Intelligence for Industries (AI4I), 2020, pp. 69-74, doi: 10.1109/AI4I49448.2020.00023.

[9]. Yoo J-H, Park Y-K, Han S-S. Predictive Maintenance System for Wafer Transport Robot Using K-Means Algorithm and Neural Network Model. Electronics. 2022; 11(9):1324. https://doi.org/10.3390/electronics11091324