# Comparative Study of decision tree and k-nn for machine tool predictive maintenance strategies

**Parth Ghube[1], Anurag Jadhav[1], Niraj Kurane[1], Om Mali[1], Rohan Pawar[1], Neha Dhere[1]**
Dr. Ganesh Dongre[2]
[1]Research Scholar, Department of Mechanical Engineering VIT Pune
[2]Dean R&D VIT Pune
parth.ghube21@vit.edu , anurag.jadhav21@vit.edu , om.mali21@vit.edu , rohan.pawar21@vit.edu , neha.dhere21@vit.edu
ganesh.dongre@vit.edu

**Abstract -** Industries are on the way to uniting today on the outlook of major technological revolution. Every day, decision-making requires a vast intake of data and its customization in the manufacturing process confronting both machines and managers. One of the primary concerns in this area is getting models to accurately anticipate when different machine modules will be in need to be maintained. The potential of performing predictive maintenance helps to improve machine uptime, control, costs and quality of production, which in turn enhances efficiency. Predictive maintenance techniques aren't generally discussed in the numerous surveys and seminars on industries that concentrate mostly on addressing machine learning and data analytics methods seeking to change the procedures of production. In this context, this paper presents systematic initiatives of predictive maintenance in industry, identifying and cataloging methods, standards and its different applications. The article outlines systematic attempts for industry predictive maintenance, identifying also suggests a brand-new taxonomy to categories this research topic taking into account the requirements of predictive maintenance in industry.

**Keywords** – Predictive Maintenance, Decision Tree, Machine Learning, k-nn, Dataset.

## I. INTRODUCTION

Manufacturers often used to blow off machine failure as an undesirable part of the experience. Machine failure happens on a spectrum and many failures cannot be identified to a precise period. While some failures are evident, visible and prove equipment inoperable, others creep up on gradually, while others deplete continuously the longer, they are neglected. Hence maintaining the machines within the optimum interval of time is very important in this case.

Failure is never a sudden effect of fall but is an effect of the sequence of various abnormalities leading to each other which starts at a certain point. Hence it needs to be predicted before it leads to actual failure or any accidents, which can be figured out with the help of functional maintenance. However, the amount of maintenance ought to be optimal, as the sources of errors might range from too little to too frequent maintenance. Maintenance that is conducted too often may enable problems to go undetected by the technician, leading in a domino effect that leads to the failure, but excessive maintenance, on the other hand, infuses chaos into the system each time. When a technician opens up a piece of machinery, there is always the risk of damage, whether it's breaking a panel, dropping a screw, accidentally flapping a wire the wrong way, removing a bolt... the possibilities are endless, and the more times the equipment is touched, the more likely it is to fail.

Maintenance is a key activity in every industry as it has a significant impact on costs and reliability. It also has a profound influence on a company's ability to keep up with the competition by offering low prices, superior service and performance. Any unexpected shutdown of machinery, equipment or devices would weaken or interrupt a company's primary business perhaps resulting in hefty fines and irreparable reputational damage to the industry. In 2013, Amazon had only 49 minutes of outage which cost the corporation $4 million in lost sales.

Regular scheduled servicing, routine checks and both scheduled and emergency repairs are all included in machine maintenance. It also comprises replacement, calibration or repair of worn, broken or misaligned elements. Routine checks too are unable to accurately predict the overall problem efficiently.

## II. PROBLEM STATEMENT

Machine Predictive Maintenance Classification

Routine maintenance also referred to as preventive or periodic maintenance is a desirable thing regardless of industry or expertise, but most firms lack the solid approaches they require.

Since original predictive maintenance datasets are generally hard to acquire and especially difficult to publish, the project presents and provide a dataset which includes 10,000 data points presented as rows with 14 features in columns which are expected from a VMC machine.

## III. LITERATURE REVIEW

The article proposes a blueprint for using Machine Learning Systems order to anticipate failure in regular production plants. Despite being a case study for a cloud data storage provider, the techniques used may be extended to many different systems and industrial uses, making it an extremely adaptable case for professionals in the field.[8]

The paper provides a comprehensive overview of recent research work on process models, purposes and strategies for predictive maintenance. The further part of the paper is followed by introducing maintenance categories such as Regression Models, PM or Predictive Maintenance. Predictive Maintenance is thoroughly studied system architectures which provide a high-level overview of Predictive Maintenance. Later part of this paper goes over the primary goals of Predictive Maintenance applications.[9] A brief overview of three types of knowledge-based approaches is clicked and also further discusses the traditional Machine Learning based

approaches.[6] Decision tree is finalized as the model for the current system due to the relevancy of the data collected and the availability of highly relevant features.

## IV. MACHINE LEARNING

Current Era claims new technologies for the industries to adapt and evolve to grab excellent opportunities. A new industry revolution is fueled by organized manufacturing, research and development, connectivity, data volume, advanced devices, customization and inventory management.

Every day large amounts of datasets are generated, stored and shared in the manufacturing firms obtained from different operations. It takes a lot of effort and consumes time to analyze such a big amount of data in order to create useful findings which is impossible to do manually. New techniques for data analysis are needed in order to detect patterns, gaps in the data and relevant interpretations in order for an application domain to make predictions. However, a variety of supervised or unsupervised machine learning methods, including Logistic Regression, clustering, Logistic Regression, Random Forest, Support Vector Machine and others, can be used to generate more accurate predictions and more accurate conclusions. To achieve the expected output, the efficiency and the time required by each algorithm to perform predictive analysis can be compared. Machine learning is a game-changing technology in today's digital world. Its applications can be found in a variety of research fields including healthcare, image analysis, manufacturing and maintenance, aviation and autonomics, among others.

A machine's maintenance and monitoring automation system are essential for its efficient operation. The main issue addressed is the lack of maintenance prediction of a machine by a prediction model so that it becomes possible to maintain a machine as soon as it is needed with the help of various real time parameters like the process temperature, rotational speed, torque and many more.

## V. MACHINE MAINTENANCE STRATEGIES

Machine maintenance strategies are widely classified into three types: preventive maintenance, corrective maintenance and predictive maintenance. Preventing and corrective maintenance excludes diagnostic and prognostic activities in determining machine maintenance. Condition-based maintenance abbreviated as CBM is carried out in predictive maintenance. CBM describes the enhancement of the maintenance plan by entirely depending on sensor data generated by the engine in real-time and preventing system failures by understanding when and how the equipment should be maintained. Creating a maintenance strategy using the CBM method will necessitate high-quality data representing the current machine's situation and state.

Condition Based Maintenance data processing with machine learning can continuously analyze data patterns from the most recent data. Machine learning in predictive maintenance can undoubtedly be used to detect machine deterioration in comparison to existing parameters. However, there haven't been enough studies on both processes using real-world industrial data. Several studies from the Proceedings ICIEOM were discovered that discussed a diagnostic or prognostic process and only used data simulation.

Different machines have various types of accomplishment and work nature. Then came scheduled maintenance activities which are also a waste of both time and human resources. The preventive method also lags and fails to predict the overall maintenance timing for any machine. Finally predictive maintenance detects machine reliability through advanced analytics and data sensing.

Predictive maintenance outperforms corrective and preventive maintenance. It really can constantly monitor diagnostic and prognostic processes to help predict errors and the RUL of the equipment (Remaining Useful Life). Machine learning models in industrial equipment can analyze various data patterns and construct failure prediction models derived from real condition monitoring using real multi-sensor data and machine failure reports. The primary goal of this research is to build a diagnostic and prognostic model by adjusting optimal machine learning specifications in support vector machines and Decision Tree for classification of equipment conditions and RUL so that the manufacturer can predict future failure intervals. In addition, machine learning parameters and methods are compared to determine which model has the highest accuracy. Based on every algorithm's accuracy, Decision Tree performs better than k-NN in diagnostic and prognostic models.

Throughout every manufacturing industry, predictive type of maintenance is the key strategy to enhance capital management. To safeguard performance deteriorates of highly advanced and more expensive machines in the workplace, predictive maintenance expertise should be considered crucial. Predictive maintenance, Machine learning, and regular interval maintenance processes are commonly employed to assess the condition of business instrumentation as a result of the emergence of new businesses in the manufacturing sector.

## VI. MACHINE LEARNING MODELS

1. Decision Tree
2. Multi-layer Perceptron
3. Logistic Regression
4. Random Forest

Algorithm: Decision tree

1. *Create root training instances.*
2. *Current node => root*
3. *for i<=c*
4. *Find the information gain for each attribute, where c denotes the number of classes and p denotes the likelihood that class I will be chosen at random.*
5. *The mean entropy of the child attributes is E(T, X), in this case. Gain from Information=E(S)-E (T,X).*
6. *end for*
7. *Features with most information aspects should be sought out.*
8. *Make this feature the current node's splitting criterion.*
9. *if information gain<=0*
10. *return current node<=leaf*
11. *end if*

Based on the values of several attribute values, the decision tree categorizes the dataset into classes. The idea of attribute selection is utilized. The two most popular techniques for accomplishing attribute selection are information gain and the Gini index. The above-described information gain methodology makes use of the offered algorithm. The input layer is located at the bottom of the multilayer perceptron, followed by number of hidden layers and the output layer is located at its top. One layer's neurons are all fully interconnected with those in the following layer.

## VII. PROPOSED MODEL

- **DECISION TREE**

Decision trees are simple to understand and explain. Although non-parametric outliers do not affect the model as they do in linear regression hence decision trees are simple to use and interact with. According to various splitting data such as Gain Ratio, Info Gain, Gini Index and Gini Coefficient, some well-known algorithms are ID3, C4.5, CART and C5.0. A decision tree can manage a variety of data including redundant attributes and missing values and it has good generalization abilities. They are noise-resistant as well as provide high-performance data with minimal computational data. These primarily employ a divide and rule strategy which delivers accurate results with highly relevant features.

Because of these factors as well as the dataset's relevance and the accessibility of extremely relevant features, a Decision tree is finalized, a Decision tree is finalized as the model for our project.

Machine learning for classification extrapolates the relationship between several input variables and the output variables or goal variables. The target variable often has two or more discrete values. There are two types of input variables: discrete and continuous. Techniques for classifying data using decision trees are a subset of such techniques. When researching the history of decision tree-based categorization methods, Quinlan (1986), the author of the seminal study "Induction of Decision Trees," concluded that the ID3 algorithm had a significant influence on the development of contemporary versions. Quinlan (1986) used the framework developed by Carbonell (1983) to divide machine learning approaches into three categories:

(1) The fundamental learning techniques used
(2) The system's representation of information obtained; and
(3) The underlying learning strategies employed

Binary partitioning refers to the process of classifying the observations in two subsets. The algorithms then move ahead in a recursive manner, dividing the two subsets further by optimizing the same metric. When the metric cannot be improved significantly further, recursive partitioning is terminated. To avoid over fitting the data, the decision trees are further trimmed, allowing the outcomes to be more broadly applicable.

The Decision Tree is a non-parametric supervised approach for regression and classification. The goal of a decision tree is to identify the values of a data point, such as its label or class, by learning straightforward decision rules inferred from data attributes. Several branches, one root, numerous interval nodes and leaves make up a decision tree. Every connection between the root and leaf nodes reflects a classification with various system circumstances or components. For classification or regression, each leaf node stands for a response or a class label. To build a Decision Tree model, first choose the most significant input variables or parameters. Then, based on the status of these variables, divide instances at the root node and subsequent internal nodes into two or more groups. The C4.5 algorithm is a popular method for generating decision trees. Machine Learning techniques based on Decision Trees have been widely used in Predictive Maintenance. First, because of the Decision Tree's nature, much effort is expended in trying to identify or categorizing the state of the real-world system. Benkercha et al., for example, introduce a new method based on the Decision Tree algorithm for detecting and diagnosing faults in grid-connected photovoltaic systems (GCPVS).

## VIII. EXPERIMENTS AND RESULTS

- **PERFORMANCE MEASUREMENT**

1. Accuracy

It is the most commonly used metric for assessing the performance of algorithms. It is defined as the ratio of accurate predictions to all other predictions. The following formula can be used to determine the correctness of the confusion matrix:

$$\text{Accuracy} = \frac{TP+TN}{TP+FP+FN+TN}$$

2. Precision

In retrieval of information, precision is defined as the number of accurate predictions. The confusion matrix can determine precision using the following equation:

$$\text{Precision} = \frac{TP}{TP+FP}$$

3. Recall or Sensitivity

The number of positives given by the model is known as recall. The confusion matrix can accurately determine recall using the formula given:

$$\text{Recall} = \frac{TP}{TP+FN}$$

4. F1 Score

The F1 score is the mean of recall and precision. To put it another way, it is the weighted mean of recall and precision. The ideal value of F1 score is 1 and its worst value is 0. The confusion matrix can accurately determine F1score using the formula given:

$$\text{F1- Score} = 2 \times \frac{Precision \times Recall}{Precision + Recall} = \frac{2TP}{2TP + FP + FN}$$

## IX. DATASET

The dataset used appears to contain 10000 data points organized as rows and 14 features organized as columns.

1. **UID**: It is Unique id which is ranging from 1 to 10000 words
2. **Product ID**: The product ID comprises of the letters H, L, or M for low i.e. 50 percent of all products, medium for 30 percent, and high for 20 percent of product quality varieties as well as a variant-specific serial number.
3. **Air Temperature [°K]**: Air Temperature is generated using a random walk process which is further normalized to standard deviation of 2 K around 300°K
4. **Process Temperature [°K]:** In addition to the air temperature +10°K, the process temperature is obtained through a [RWP] random walk process and standardized to a 1°K standard deviation.
5. **Rotational Speed [rpm]**: Calculated rotational speed using 2860 W of power layered with normally distributed noise.
6. **Torque [Nm]**: The typical distribution of torque values is 40 N-m, with a 10 N-m and there are no negative values.
7. **Tool Wear [min]**: The top-tier versions H, M, and L increase tool wear by 5/3/2 minutes respectively during the process.
8. The equipment's failure at a specific data point for either of the failure modes is indicated by the label "**Machine failure**."

## X. SYSTEM USED

**Google Colab**

Why Google Colab?

Google Colab is an extremely good tool for performing deep learning tasks. It is a free hosted 'Jupyter' notebook that doesn't even demand any setup and provides free access to Google computing resources such as GPUs (Graphical Processing Unit) and TPUs (Tensor Processing Unit).

## XI. CONFUSION MATRIX

A technique used to assess the effectiveness of the classification models for a certain set of test data is the confusion matrix. It is basically a way to represent the result of the model in an efficient way and to judge them. It can only be determined if the true values for test data are known. The matrix can solely be easily understood, but the connected terminologies may be confusing as the relation between predicted and actual values must be properly understood. As it shows the errors in the model performance in the form of a matrix hence additionally it is also known as an Error Matrix.

It is arguably the simplest method for assessing the effectiveness of a classification problem with many output classes. As seen below, a confusion matrix is a table with 2 dimensions, Actual values and Predicted values with four columns: True Negatives (TN), True Positives (TP), False Negatives (FN) and False Positives (FP)

The definitions of the terms used in the confusion matrix are as follows:

1) True Positives [TP] -In this instance, the actual and projected classes are both 1.
2) True Negatives [TN] - In this case, both the actual and the predicted class of the data point are 0.
3) False Positives [FP] - In this case, the actual class is 0 and the predicted class's data point is 1.
4) False Negatives [FN] − In this case the actual class of data point is 1 and the predicted class's data point is 0.

## XII. DECISION TREE VS K-NN

1) Confusion Matrix

**a) Decision Tree**

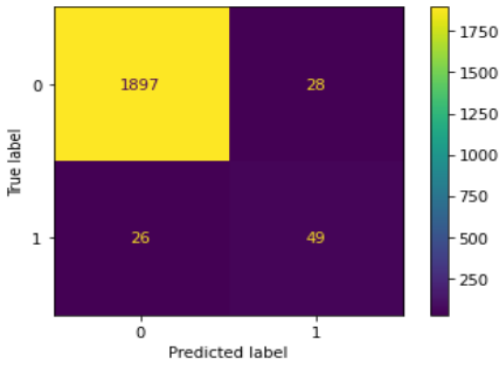


Figure no.1 confusion matrix of decision tree

From the figure above:
True Positive (TP): 1897

It states that our prediction model has predicted that the machine will not fail 1897 times which is predicted correctly and efficiently followed by the actual model.

False Positives (FP): 28

It states that our prediction model has predicted that the machine will not fail 28 times but it is incorrectly predicted as it does not fail in actuality.

False Negatives (FN): 26

It states that our prediction model has predicted that the machine will not fail 26 times but it is incorrectly predicted as the machine fails in actuality.

True Negatives (TN): 49

It states that our prediction model has predicted that the machine will fail 49 times which is predicted correctly and efficiently followed by the actual model.

**b) k-NN**

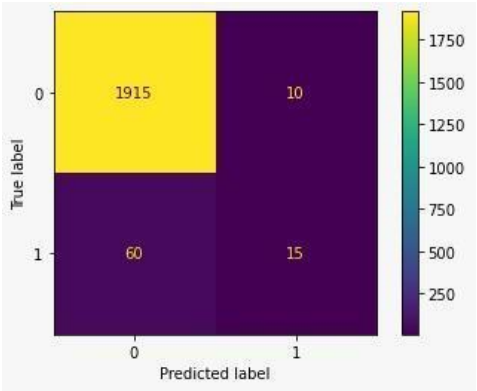


Figure no.2 confusion matrix of k-nn

From the figure:
True Positive (TP): 1915

It states that our prediction model has predicted that the machine will not fail 1915 times which is predicted correctly and efficiently followed by the actual model.

False Positives (FP): 10

It states that our prediction model has predicted that the machine will not fail 10 times but it is incorrectly predicted as it does not fail in actuality.

False Negatives (FN): 60

It states that our prediction model has predicted that the machine will not fail 60 times but it is incorrectly predicted as the machine fails in actuality.

True Negatives (TN): 15

It states that our prediction model has predicted that the machine will fail 15 times which is predicted correctly and efficiently followed by the actual model.

**2) Observation Table**

| Model Type | Accuracy | Precision | Recall | F1 score |
|---|---|---|---|---|
| **Decision Tree** | 0.973 | 0.985 | 0.986 | 0.986 |
| **k-NN** | 0.965 | 0.995 | 0.970 | 0.982 |

**3) Graph Decision Tree Vs KNN**

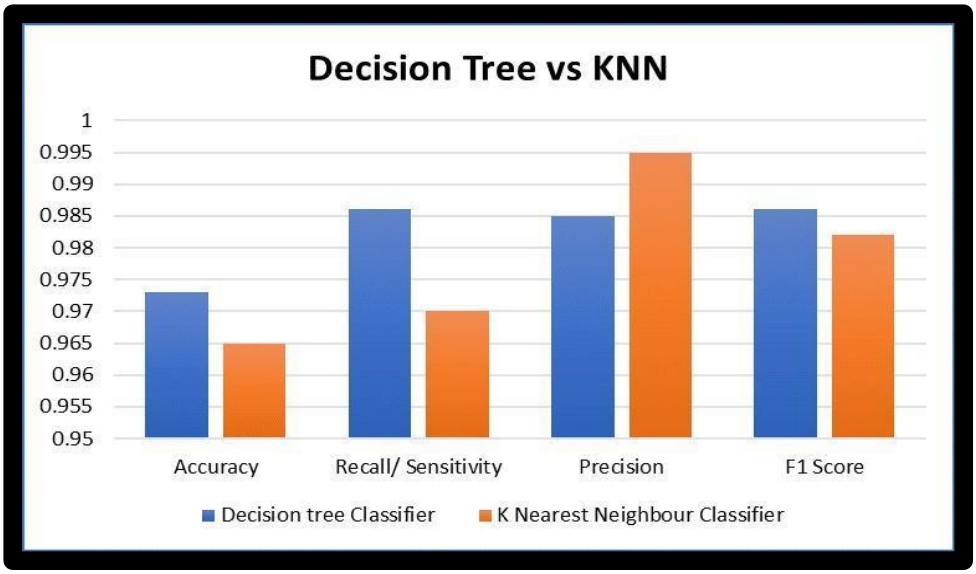


Figure no.3 decision tree vs k-nn

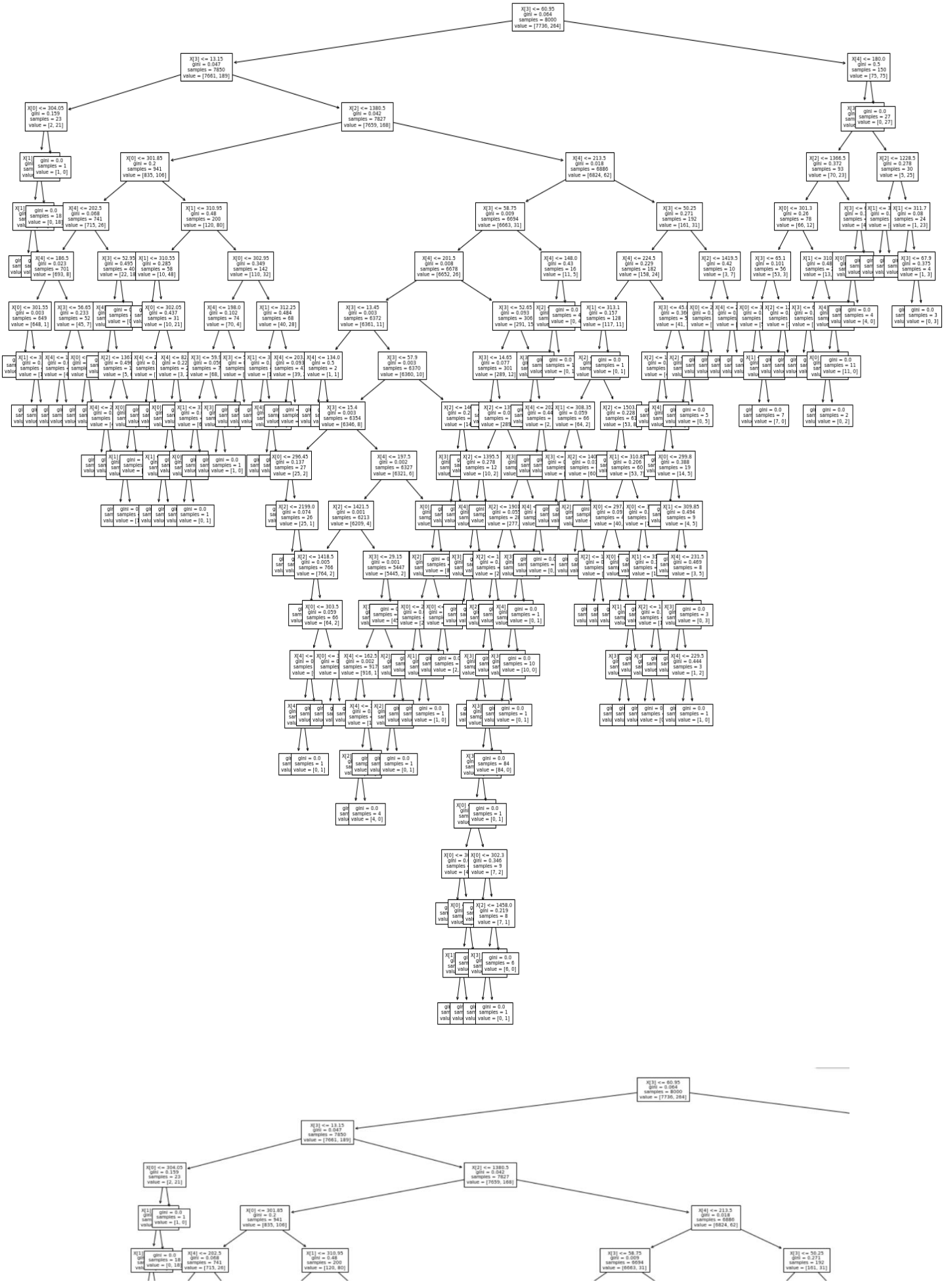

Figure no.4 Decision tree

## XIII. CONCLUSION

Maintenance activity requires the different strategies and planning in order to satisfy the requirements of productivity, quality and safety. Expanding industry conceptions bring with them new opportunities and difficulties. Predictive maintenance was recognized as the subject of our study, in which the aim was to comprehend the methodologies employed and the applications already in use. 15 papers addressing techniques, architectures and technologies relevant to the use of predictive maintenance were found using a systematic literature review process. A critical evaluation of the studies was carried in light of the key issues in the industry, such as the difficulty in visualizing possible standardizations and their implementation.

The study addresses scholars and practitioners with expertise in machine learning and the maintenance field in an effort to create the foundation for multidisciplinary interactions. A significant contribution in the identification of open problems and the research areas which may aid researchers in identifying open research issues. There are a number of significant conclusions that can be made:
(1) Research activity would rise if there were more publically available data for predictive maintenance systems. Using data from other sources can also raise accuracy and open up new applications.
(2) Drivetrain transformation could benefit from ML-based Predictive Maintenance techniques. The use of deep learning techniques in predictive maintenance is anticipated to go even further, but this will necessitate the adoption of approaches that are specifically customized for efficiency, interpretability and data accessibility.

## XIV. REFERENCES

[1] A Comparative Study Of State-Of-The-Art Machine Learning Algorithms For Predictive Maintenance
[2] A Survey Of Predictive Maintenance: Systems Purposes And Approaches
[3] A Systematic Literature Review Of Machine Learning Methods Applied To Predictive Maintenance
[4] Designing Predictive Maintenance Systems Using Decision Tree-Based Machine Learning Techniques
[5] Machine Learning Approach To Predictive Maintenance In Manufacturing Industry - A Comparative Study
[6] Predicting Service Industry Performance Using Decision Tree Analysis
[7] PREDICTIVE MAINTENANCE AND MONITORING OF INDUSTRIAL MACHINE USING MACHINE LEARNING
[8] Predictive Maintenance Enabled By Machine Learning: Use Cases And Challenges In The Automotive Industry
[9] Predictive Maintenance In The Industry 4.0: A Systematic Literature Review.
[10]Predictive Maintenance Using Machine Learning Based Classification Models