COBEM 2023 27TH INTERNATIONAL CONGRESS OF MECHANICAL ENGINEERING

ABCM Brazilian Society of Mechanical Sciences and Engineering
27th ABCM International Congress of Mechanical Engineering
December 04-08, 2023, Florianópolis-SC, Brazil

FAULT DETECTION IN WIND TURBINES WITH TEMPERATURES ANALYSIS AND STATISTICAL MODELS.

First Author's Name Paiva, Lucas Lira de.
Second Author's Name Leite, Gustavo de Novaes Pires.
Third Author's Name Ochoa, Alvaro Antonio.
Institution and address for first and second authors - if the same: Federal Institute of Technology of Pernambuco, Av. Prof Luiz Freire, 500, Recife/PE – CEP 50740-545
e-mails: llp1@discente.ifpe.edu.br, gustavonovaes@recife.ifpe.edu.br, ochoaalvaro@recife.ifpe.edu.br

Fourth Author's Name Costa, Alexandre.
Fifth Author's Name Petribú, Leonardo.
Institution and address for fourth and fifth authors – if the same: CER-UFPE – Center for Renewable Energy of the Federal University of Pernambuco, Brazil.
e-mails: alexandre.acosta@ufpe.br, leonardo.brennand@ufpe.br

Sixth Author's Name Souza, Marrison Gabriel Guedes de.
Institution and address for third author: NEOG – New Energy Options Geração de energia, Brazil.
e-mail: marrison.souza@neog.com.br

Abstract. *In a world where the energy demand is constantly increasing, wind power is essential because of its sustainable nature, competitive costs, and outstanding potential to produce energy. However, wind turbines are giant machines with high installation, operation, and maintenance costs. A wrong maintenance strategy could ruin the cash flow of a wind farm, compromising the investment of decades. Wind turbines have monitoring systems for evaluating whether they work within operational limits. These systems store a massive amount of valuable data, which could provide an early indication of faults occurring in the machines and, consequently, avoid unexpected maintenance costs. Temperature is commonly used to monitor the condition of mechanical machinery. The wind turbine supervisory system monitors critical components' temperature and stores this information in a database. In this sense, the present work proposes developing a machine-learning model based on the analysis of the temperature of wind turbine components to determine, in advance, whether it is operating in its normal behavior. The present methodology proposes implementing normal-behavior models using different machine learning algorithms (artificial neural network, random forest, and k-nearest neighbor) to detect faults in wind turbine components such as the gearbox, main bearing, and generator. Operational data from wind turbines installed in Brazil are used, and challenges about using unlabeled real data are discussed throughout the paper. Some challenges are filtering data, selecting variables and data windows for training, and validating the models. The models can identify deviation from normal behavior, characterizing a fault, even before the supervisory system triggers the alarms. Results also present what methods perform efficiently and if there are differences regarding the analyzed component. Anticipated actions from the maintenance staff to correct the faults can be carried out in a planned and efficient way, which not only preserves the wind turbine but also increases the wind farm's key performance indicators. Different detection periods were identified depending on each component's dynamics and the model's particularities. The better-performing models were artificial neural networks and decision trees, detecting faults from 80 to 100 days in advance for the gearbox and 90 to 120 days before for the main bearing.*

Keywords: *Temperature Analysis, Wind turbine, Statistical models.*

1. INTRODUCTION

With the increasing demand for electrical energy, renewable energy generation devices are receiving more attention due to the need for their implementation. This is because energy production from polluting and depletable sources such as oil, natural gas, and coal has proven to be harmful to the environment, for example, contributing to the increase in the greenhouse effect. Wind turbines are highly sustainable options, but there are significant maintenance expenses that could be mitigated if failures in wind turbines could be "predicted" at an early stage. With reduced maintenance costs, greater investment in renewable energy worldwide could be possible, as it is a cheaper and more environmentally friendly energy source. Predictive maintenance is based on the machine's behavior and knowing the exact moment to perform strategic shutdowns, minimizing costs of corrective and preventive maintenance that would not be necessary. There are various ways to "predict" such failures, and one of them is by observing and analyzing the behavior of a temperature variable that is directly linked to a major component of a wind turbine (e.g., a component with high replacement/repair costs; a component with high downtime periods after a failure; a component with a high rate of failures).

The wind energy capacity in Brazil has been increasing every year and is expected to grow exponentially to replace fossil fuels with renewable energy. Wind turbines are crucial for converting wind energy into electrical energy. However, these wind turbines are subject to operational challenges and failures that can lead to reduced performance and costly repairs. Early detection of failures is crucial to ensure optimal operation and longevity of wind turbines. This paper proposes an approach using temperature analysis related to the major components of the wind turbine and statistical models. For example, monitoring the main bearing, gearbox, and generator temperatures can detect deviations from normal operating conditions, allowing for predictive maintenance actions and minimizing downtime.

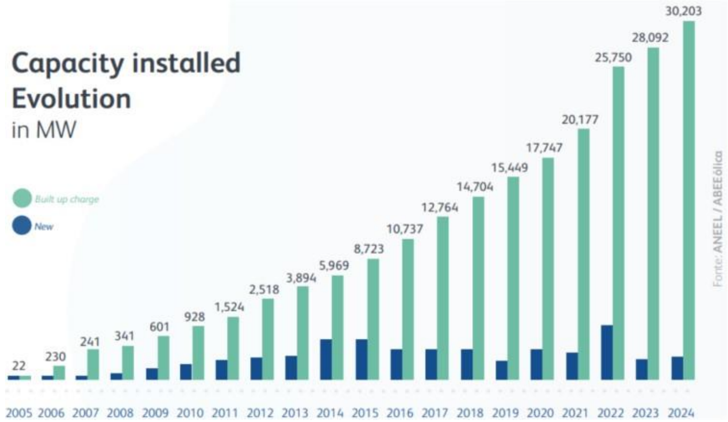


Figure 1: Wind Energy Capacity in Brazil over the years. Source: Airswift, 2023.

1.1 Machine learning

Machine learning is a subfield of artificial intelligence that focuses on the development of algorithms and statistical models that enable computational systems to learn patterns and make decisions based on the provided training data. The goal is to empower machines to learn patterns and make decisions or predictions based on those patterns, improving their performance over time. In this paper, the Python programming language was used, which is easily accessible and understandable, with rich libraries for plotting graphs that were of great importance for the project development. Machine Learning has many applications in different fields, including classification and regression, which were the applications used in this study.

I. Classification: Machine Learning can be used to classify data into different categories based on specific features or attributes. For example, it can classify emails as spam or non-spam, identify whether a financial transaction is fraudulent or not, or even classify images into different categories.

II. Regression: Machine Learning can also be used for regression analysis, where the goal is to predict a continuous value based on a set of input variables. This can be applied to predict real estate prices based on their characteristics, forecast future demand for a product based on historical sales data, among others.

These are just a few of the many applications of Machine Learning. The breadth and versatility of this field make it possible to apply it in various domains, such as medicine, finance, manufacturing, agriculture, transportation, among others. The advancement of Machine Learning has also propelled research in artificial intelligence and enabled the development of more intelligent and autonomous systems.

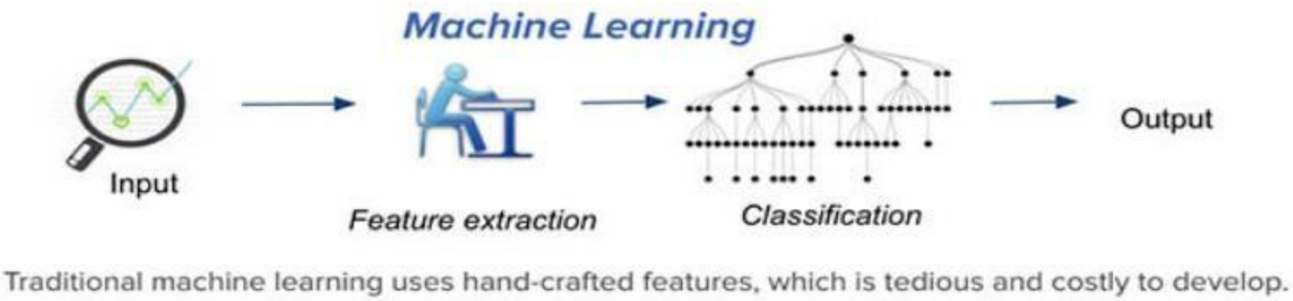

Machine Learning
Input
Feature extraction
Classification
Output
Traditional machine learning uses hand-crafted features, which is tedious and costly to develop.

Figure 3: Simple description of the machine learning process. Source: Merkle, 2023.

1.2 Literature review

As previously mentioned, the demand for electrical energy worldwide has been increasing over time. However, the environmental impacts are now a significant concern and should be avoided by all means. The use of renewable energy sources is becoming increasingly attractive to everyone, and wind energy is a successful alternative for countries with favorable climatic conditions. In Brazil, wind energy can be widely utilized due to its ideal climatic conditions, as demonstrated in the study [01]. However, there are challenges to be faced, such as the high cost of equipment and maintenance, as seen in [02]. Therefore, there is a need to minimize unnecessary expenses, particularly in terms of preventive maintenance. The following study presents a way to reduce and replace such preventive maintenance costs with predictive maintenance, which is more cost-effective.

As extensively discussed in [03], the selection of the most important features for diagnosing faults in the large component under study depends on the data sampling time available and the specific large component being investigated. In this case, the focus was solely on the gearbox, whereas [04] discusses the same but with a different large component, the generator.

That being said, this paper is not limited to a specific large component but rather focuses on multiple large components of the wind turbine, with the analysis being determined by temperature variables directly related to these components. Zaher et al. [05] explore fault diagnosis in wind turbines through temperature data analysis. The study focuses on developing a model for normal behavior of gearbox oil temperature, gearbox bearing temperature, and generator winding temperature using artificial neural networks (ANN). The model takes into account input variables such as power generation, ambient temperature, and historical values of target temperatures at one and two previous time instances. The researchers aimed to leverage temperature patterns to enhance the understanding and detection of potential faults in wind turbines using this approach. This paper uses the same input variables of temperature data but addresses other computational models, such as K-nearest neighbors (KNN) and Random Forest (RF), which is a feature utilized in [06], where it is based on a linear regression model for analyzing temperature variability in wind turbines.

Thus, the implementation of statistical models becomes necessary to perform graphical analysis when the computational model is in operation and to select the best methods for application in wind turbines. This is based on a study that compares threshold determination methods, as shown in [07]. Once it was chosen that the Tukey method was the threshold determination method, the process described in [08] was initiated, which presents statistical techniques for determining not only thresholds but also statistical relative entropies. Furthermore, it establishes a relationship and applies these techniques to collected data, using the SCADA system in the case of this paper, similar to [09]. In [09], data collected by the SCADA system from healthy turbines is utilized, and the Tukey method is applied with the aim of improving the accuracy of the power generation characteristic curve in wind turbines.

1.3 Paper structure

Section 2 describes the methodology used in this study, including the statistical and computational models employed for data analysis, the methodology itself, the input data, and the discussion on the theory and utilization of the computational models used in this paper.

Section 3 discusses the results obtained from the analysis of each statistical and computational model. The model's behavior for different large components is also compared, along with the discussion about the performance of each model for the studied large component.

Section 4 presents the study's final conclusions, including discussions and applications of the examined model.

Finally, at the end of the paper, the references of the studies used as a basis and guidance for developing the presented paper are provided.

2. METHODOLOGY, MODELS AND INPUT DATA

The proposed methodology in this paper is based on fault detection in wind turbines, involving the following steps: data collection, preprocessing, feature extraction, statistical modeling, and fault detection. Temperature data from the main bearing, generator, and gearbox are collected through sensors embedded in the wind turbine system. Preprocessing techniques are applied to remove noise and outliers from the data. Feature extraction algorithms are then used to derive relevant features from the time series temperature data. Statistical models such as regression analysis, neural networks, random forests, and k-nearest neighbors are developed using the extracted features. Finally, fault detection is performed by comparing the observed temperature patterns with the expected normal behavior of the components.

2.1 Methodology

In Figure 4, a flowchart depicting the stages of the proposed objective can be observed based on the behavior of the temperature variables that will be analyzed.

In the first point, the behavioral change of the temperature variable being analyzed by the model is mentioned, where the variables are influenced by external factors that cause them to behave differently over time. Next, in the second point, the detection of this change in temperature behavior is addressed, where statistical and computational models allow visualizing this change in the behavior of the temperature variables. The topic of temperature behavior analysis is then discussed, where despite the behavior change, a thorough analysis is necessary to determine whether this alteration in variable performance indicates an imminent fault or if it is due to an emerging external factor that will not result in a failure. Based on this analysis, the decision of intervention or non-intervention (depending on the outcome of the previous analysis) is made by maintenance personnel to repair the specific item that is not functioning as expected.

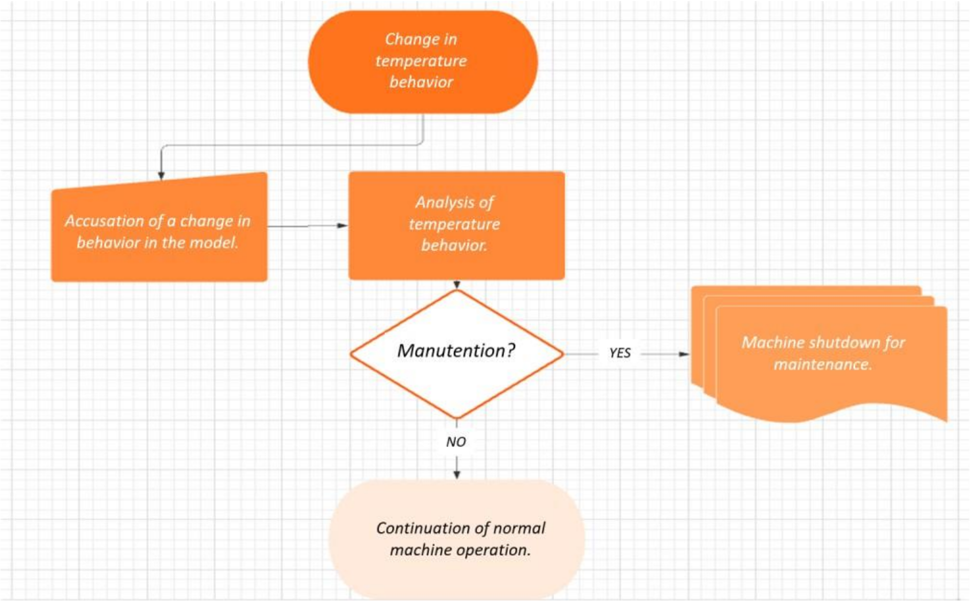

Change in temperature behavior
Accusation of a change in behavior in the model.
Analysis of temperature behavior.
Manutention?
YES
Machine shutdown for maintenance.
NO
Continuation of normal machine operation.

Figure 4: Flowchart of the proposed objective methodology. Source: Author, 2023.

2.2 Models and their methodology

Figure 5 shows a flowchart that consists of 8 main steps, and each of these steps involves different areas of knowledge for model creation, ranging from statistics and programming to the analysis of mechanical thermodynamic engineering itself.

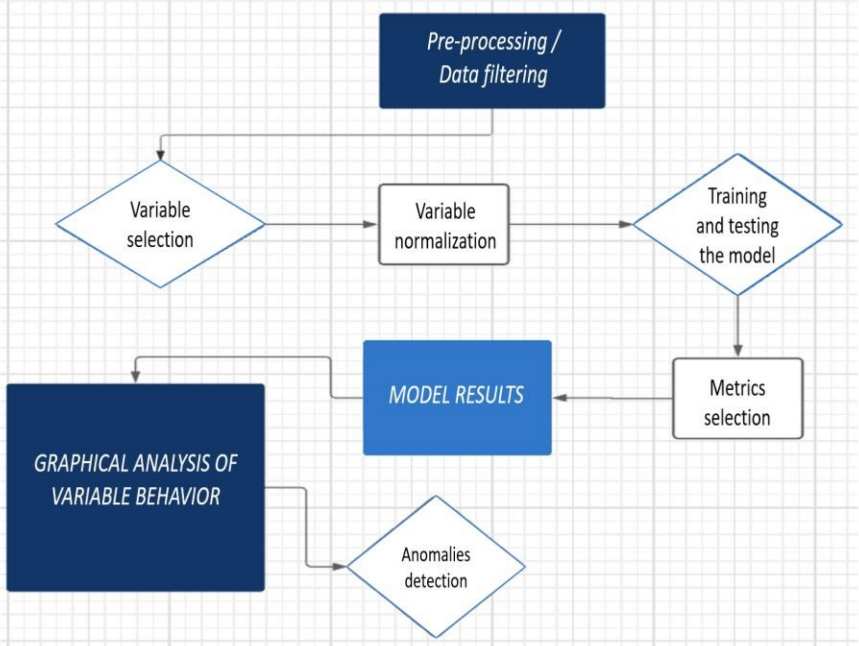

Pre-processing / Data filtering
Variable selection
Variable normalization
Training and testing the model
Metrics selection
MODEL RESULTS
GRAPHICAL ANALYSIS OF VARIABLE BEHAVIOR
Anomalies detection

Figure 5: Flowchart of the computational model methodology. Source: Author, 2023.

2.2.1 Scada System

SCADA (Supervisory Control and Data Acquisition) is a control and data acquisition system used in various industrial sectors, including wind power generation. In the context of wind turbines, the SCADA system is responsible for collecting, monitoring, and remotely controlling the turbines' operations. The SCADA system in wind turbines collects a wide range of real-time information, which is exemplified in the table below:

SCADA data collection
Gearbox Oil temperature
Main Bearing temperature
Generator Bearing temperature
Wind speed
Wind direction
Generated Power
Rotor speed
Ambient Temperature

Table 1: Examples of Sensors present in the SCADA system.

For the present study, SCADA data is collected every 10 minutes of wind turbine operation for 10 years. The data is averaged, and the maximum and minimum values are recorded for each 10-minute interval. These averages are documented in reports and utilized in the current computational model.

Furthermore, the SCADA system also enables operators to remotely adjust and configure operational parameters of the wind turbine, such as rotational speed and pitch control of the blades, among other control parameters. This enables wind turbine performance optimization and enhances maintenance efficiency and necessary interventions.

In summary, the applications of the SCADA system for wind turbines are diverse and include:

I. Performance monitoring: The SCADA system enables continuous turbine performance monitoring by providing information on energy production, wind speed, vibration, and other parameters. This helps identify potential faults.

II. Remote control: Operators can remotely control the turbines through the SCADA system. This includes adjusting the blades' rotational speed, controlling the blades' pitch angle to optimize energy capture, and performing other control operations.

III. Fault diagnosis: The SCADA system records historical and real-time data, allowing for the detection of faults or abnormal behavior in the turbines. This enables more efficient preventive or corrective maintenance, reducing downtime and associated costs.

IV. Data analysis: The SCADA system provides significant data, including temperature data. Analyzing this data can reveal patterns, trends, and relationships between variables. For example, temperature analysis can help identify anomalies such as component overheating or variations outside the expected range.

2.2.2 Data filtering and preprocessing

Figure 5 identifies the first step as the data preprocessing and filtering stage. In this step, various filters are applied to the raw data collected from the SCADA system to eliminate data on the verge of failure or in a failed state. The goal is to remove data that does not exhibit the ideal behavior expected for the wind turbine in its normal operation, and only insert healthy data into the computational model. In Figures 6 and 7, it is possible to visualize the power curve (power generation variable of the wind turbine) without preprocessing and with preprocessing. In Figure 6, the curve shows scattered data points and the wind turbine operating with a power limitation. Although the data continues to be collected, it is understood that they should be eliminated for better analysis and model training. In Figure 7, the power curve represents the "ideal" behavior where the turbine operates as expected.

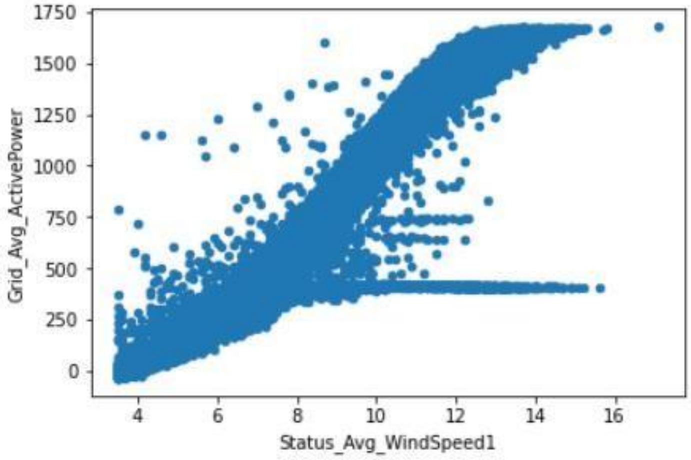


Figure 6: Power curve graph of the wind turbine without data preprocessing and filtering (unhealthy data). Source: Author, 2023.

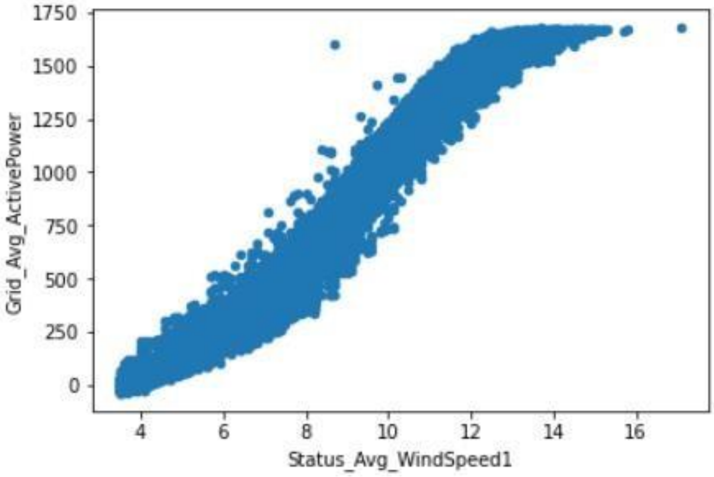


Figure 7: Power curve graph of the wind turbine with data preprocessing and filtering (healthy data). Source: Author, 2023.

The applied filters are functions in the Python programming language, where functions are defined to identify points considered outliers through Binned data analysis[10]. Unfortunately, the SCADA system (data collection system) cannot identify changes in power generation behavior in a wind turbine. These changes can occur due to power limitations in the wind turbine, such as when a faulty component restricts power generation to preserve the turbine’s health. The system only detects alarms but continues turbine operation and data collection. However, this data is considered “unhealthy” as the wind turbine is not operating at 100% capacity. Therefore, an analysis technique is needed to identify these power limitations (as shown in Figure 6) and the main objective is to extract the “unhealthy” data by cross-referencing alarm reports with the turbine’s failure and maintenance history. This ensures that the computational model is trained only with “healthy” data.. Additionally, wind speed data outside the operational limits for power generation are disregarded. Since there are two wind speed sensors in the system, the difference between the wind speeds has limits, meaning the speeds cannot be significantly discrepant.

2.3 Input data

2.3.1 Variable selection

For this analysis, the model relied on turbines that exhibited failures or anomalies in the major observed components during their operation within these 10 years of data collected by the SCADA system. For the selection of input variables, evaluations were made to determine which variables would be selected depending on the major component being analyzed. Correlation matrices aided in this analysis, as shown in Figure 9. In the gearbox case, the selected variables have the strongest relationships with this major component, as presented in Tables 2, 3, and 4. In addition to selecting the variables, it is necessary to select the wind turbines that experienced failures in the major components under study, many wind turbines were analyzed, each with the major component under study exhibiting anomalies.

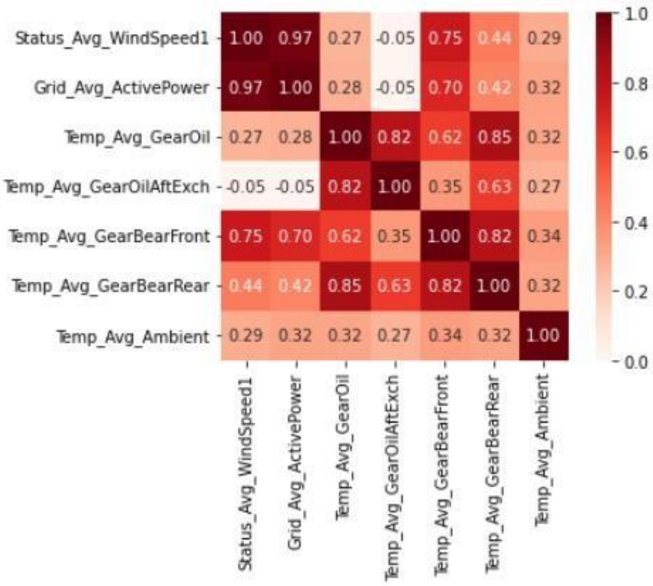


Figure 9: Correlation matrix of input variables for the gearbox. Source: Authors, 2023.

Gearbox - Target variables = gearbox oil temperatures
Wind speed
Active power generated
Gearbox oil temperature
Gearbox oil temperature after heat exchanger
Gearbox front bearing temperature
Gearbox rear bearing temperature
Ambient temperature

Table 2: Input variables for the major component - Gearbox.

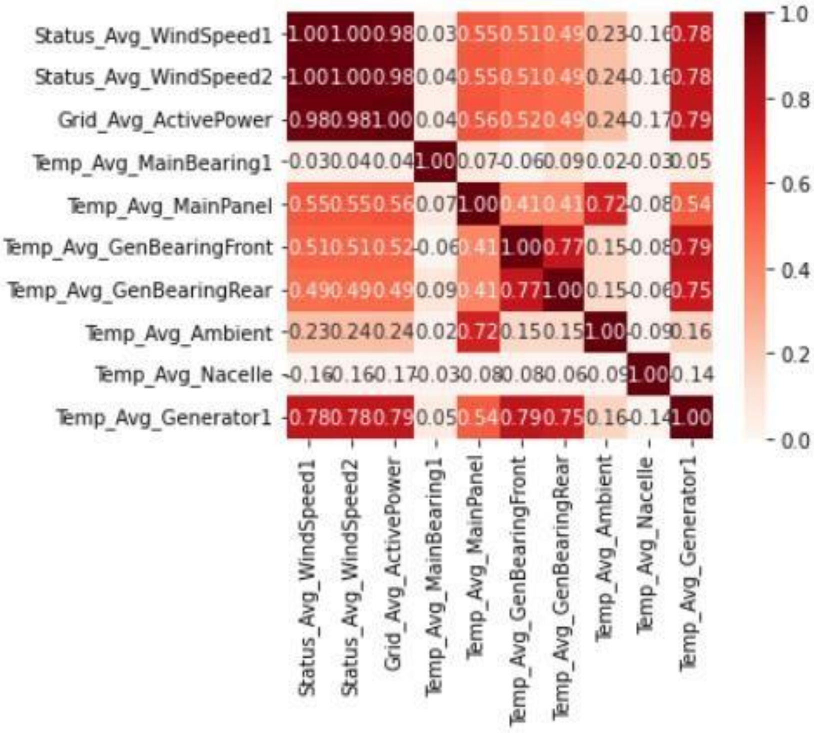


Figure 10: Correlation matrix of input variables for the main bearing. Source: Authors, 2023.

Main Bearing - Target variable = Main Bearing temperature
Wind speed 1 and 2
Active power generated
Main Bearing temperature
Main control panel temperature
Generator Front bearing temperature
Generator Rear bearing temperature
Nacelle temperature
Ambient temperature

Table 3: Input variables for the main bearing.

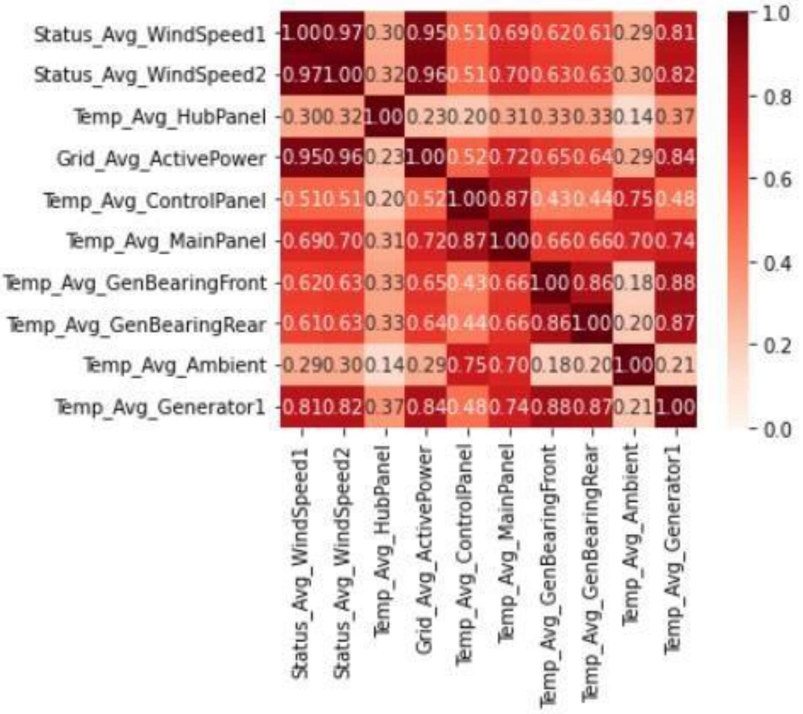


Figure 11: Correlation matrix of input variables for the generator. Source: Author, 2023.

GENERATOR - Target Variable = Front generator bearing temperature
Wind speed 1 and 2
Active power generated
Control panel temperature
Main panel temperature
Generator Front bearing temperature
Generator Rear bearing temperature
Generator temperature
Ambient temperature

Table 4: Input Variables for the major component - Generator.

2.3.2 Variable normalization

The normalization of variables is done to standardize the scale of the input variables before feeding them into a computational model for data analysis. This step is important because many machine learning algorithms and models are sensitive to the scale of variables. Normalization helps prevent variables with widely different scales from disproportionately impacting the analysis or model.

Some reasons why variable normalization is performed are the (I) weight balance, where machine learning algorithms that use distance or similarity measures between variables, such as linear regression, k-means, SVM (Support Vector Machines), and neural networks, can be influenced by the scale of the variables. A variable with a much larger scale than others can dominate the modeling process and result in biased weights or coefficients. Normalization helps balance the influence of each variable by assigning them a comparable scale; (II) faster convergence, here, the optimization algorithms used in machine learning models, such as gradient descent, converge more quickly when variables are on the same scale. Normalizing the variables accelerates the optimization process by reducing the number of iterations required to reach a satisfactory result (III) reduced impact of outliers, in this expression, outliers in a variable can significantly affect certain algorithms. Normalizing the variables can reduce the impact of these outliers, making the model more robust and less sensitive to extreme values (IV) interpretation of coefficients, where in some cases, normalizing the variables also facilitates the interpretation of coefficients or weights assigned to each variable. When variables are on the same scale, it is easier to compare the magnitude of coefficients and understand their relative importance for the model (V) preservation of privacy, in some scenarios where data privacy is a concern, normalization can be applied as an anonymization technique, where original values are replaced with normalized values that conceal sensitive information.

In summary, variable normalization is an important step in data analysis and machine learning modeling as it helps improve the performance, stability, and interpretability of models and facilitates comparison between variables. For this task, the "StandardScaler" or "Z-score normalization" was used. This technique is widely used in data analysis and machine learning modeling to normalize variables to a scale with zero mean and standard deviation of one.

The formula for standardizing a variable \`X\` is given by:

$$Z = \frac{X - \mu}{\sigma}$$

Where:
- \`Z\` is the standardized value of variable \`X\`,
- \`X\` is the original value of the variable,
- \`μ\` is the mean of the variable,
- \`σ\` is the standard deviation of the variable.

The \`StandardScaler\` function implements this standardization by calculating the mean and standard deviation of the training data and then applying the transformation to normalize the data. When fitting the \`StandardScaler\` to the training data using the \`fit\` method, it calculates the mean (\`μ\`) and standard deviation (\`σ\`) of each variable in the training data. These statistics are necessary for the subsequent standardization. The transformation is then applied to the training data using the \`transform\` method. In this process, each value \`X\` is subtracted by the mean \`μ\` of the corresponding variable, and the result is divided by the standard deviation \`σ\`. This ensures that the resulting data has a mean of zero and a standard deviation of one. Normalizing the variables is an important step in data analysis and machine learning modeling as it helps to avoid issues related to variable scales. By standardizing the variables, they can be directly compared, improving the stability of the models and facilitating the interpretation of coefficients or weights assigned to each variable. The \`StandardScaler\` function from the \`sklearn\` library simplifies the normalization process, allowing data scientists and researchers to easily apply this technique to their datasets, contributing to a more robust and reliable analysis.

2.3.3 Training and testing the model

A linear regression model was used to train and test the model. In machine learning, linear regression is widely employed as a supervised learning technique to predict continuous numerical values based on a set of input variables. Linear regression aims to find a mathematical equation that represents the linear relationship between the independent variables and the dependent variable. This equation is used to make predictions or estimates of the dependent variable's values based on the independent variables' known values. The equation for linear regression is represented by:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_\mathrm{n} X_\mathrm{n}$$

Where Y is the dependent variable, $X_1, X_2, ..., X_n$ are the independent variables, $\beta_0, \beta_1, \beta_2, ..., \beta_n$ are the regression coefficients representing the slope and intercept of the regression line. Linear regression has various applications in different areas.

Based on the above topics, linear regression can be applied to problems such as price prediction, trend analysis, impact assessment of variables, and modeling relationships between variables. It is a versatile and widely used tool in machine learning, enabling the understanding and prediction complex phenomena based on available data.

For the study in question, as referenced in [11], 70% of the data was used for training and 30% for testing a linear regression model. Validation was performed over a period with complete and unfiltered data. The model's training, testing, and validation periods vary depending on the major component being studied and the selected wind turbine. **For exemplification, a specific turbine was selected**, which had its gearbox compromised due to a breakage in the shoulder of the gearbox bearing. For this turbine, the training period was from January 1, 2016, to June 5, 2017, and the testing period was from June 5, 2017, to December 31, 2019. Three types of machine learning algorithms were used, namely:

I. Neural network (RNe)

Machine learning aims to empower computers to learn from data, enabling them to make decisions and perform complex tasks. Neural networks are one of the most prominent machine learning methods, highly effective in classification problems, regression, image processing, and speech recognition, among others. Neural networks are composed of interconnected units called artificial neurons or processing units. These neurons are organized into layers, with the input layer responsible for receiving the data and the hidden and output layers performing the processing and producing results.

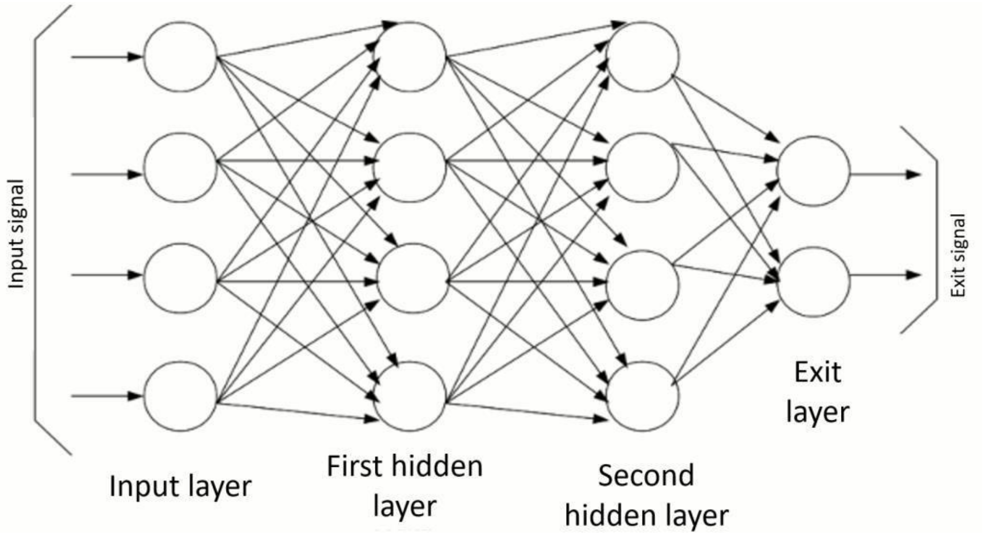


Figure 12: Layout of a Neural Network-style machine learning model. Source: Monolito Nimbus, 2017.

Each connection between neurons has a weight that determines the importance of that connection in signal transmission. There are various neural network architectures used in machine learning. One of the most common is the Feedforward Artificial Neural Network (ANN), also known as Multilayer Perceptron (MLP). In this architecture, neurons are organized into successive layers, where each neuron receives information from the neurons in the previous layer and passes its output to the neurons in the next layer.

Neural networks are a powerful method in machine learning, with the ability to handle complex and non-linear problems. However, choosing the right architecture, proper tuning of hyperparameters, and availability of high-quality training data are critical factors in achieving good results. Neural networks continue to be the subject of study and research, with ongoing advancements that expand their potential and applicability in various domains.

II. Random Forest (RFo)

Random Forest is a machine learning algorithm that combines the strength of multiple decision trees for classification and regression tasks. This technique stands out for its ability to handle complex data and highly correlated input variables. Random Forest is constructed from a set of decision trees, where each tree is trained on a random sample from the training dataset. During training, each tree is developed independently by splitting the data into different nodes based on information gain or Gini index criteria. In the classification process, Random Forest combines the predictions from each tree to determine the most likely class. In regression, the average of predictions from all trees is calculated.

This combination of predictions reduces the tendency for overfitting, making Random Forest a robust model capable of generalizing well to new data.

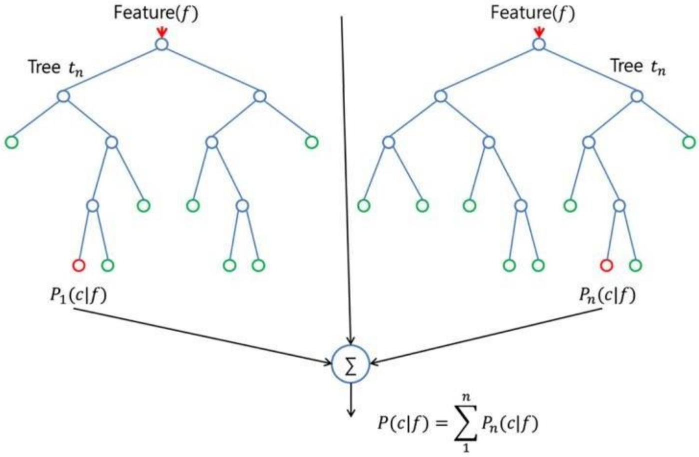


Figure 13: Layout of a Random Forest machine learning model. Source: Medium, 2018.

Random Forest is a versatile and effective algorithm that handles classification and regression problems in different domains. Its ability to handle correlated variables and its robustness against overfitting make it a popular choice in machine learning. However, adjusting its hyperparameters properly and having a representative dataset to achieve better results is important.

III. K - Nearest Neighbors (KNN)

K-Nearest Neighbors (KNN) is a supervised machine learning algorithm based on the principle that similar instances tend to belong to the same class or have similar values. This algorithm is non-parametric, which means it doesn't make specific assumptions about the data distribution. KNN determines the class or value of an unknown data point based on the classes or values of its nearest neighboring data points. The distance between points is calculated using metrics such as Euclidean distance. The parameter K defines the number of neighbors considered for decision-making. It assigns the most frequent class among the K nearest neighbors to the unknown data point. For regression, the predicted value is calculated as the average of the values of the K nearest neighbors. The value of K can vary depending on the problem and should be chosen carefully, considering the trade-off between bias and variance.

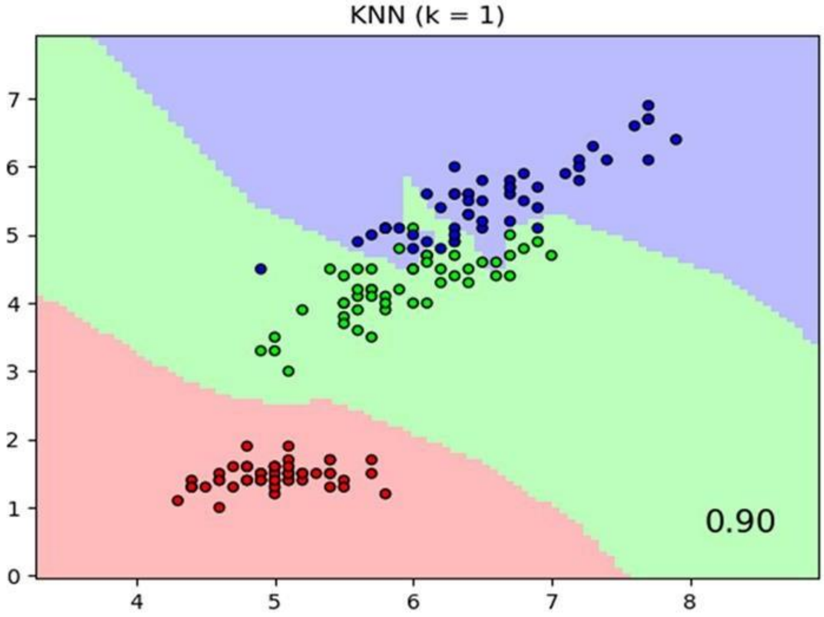


Figure 14: Layout of a K-Nearest Neighbor machine learning model. Source: MQL5, 2023.

K-Nearest Neighbor is a simple yet effective machine learning algorithm suitable for problems with well-defined structure and labeled data. However, its effectiveness can be impacted by imbalanced or high-dimensional data. It is important to consider these factors when applying the KNN algorithm and adjust the value of K according to the problem's characteristics.

2.3.4 Graphic limits determination

The definition of upper and lower limits in the residual temperature plots is a common practice to aid in the analysis of residuals and identify patterns or anomalous behavior in the data. Residual plots are used to assess the quality of the regression model fit and check for patterns in the residual errors, i.e., whether the errors exhibit any systematic structure. They are constructed by plotting the residuals (the difference between the observed values and the values predicted by the model) against the independent variables or against time, depending on the context of the problem. Defining upper and lower limits in the residual plots allows identifying regions where the residuals deviate beyond what is expected, indicating potential issues or unusual patterns in the data. These limits can be based on statistical criteria such as standard deviations of the residuals or confidence intervals. It is possible to identify points that fall outside the established limits by setting these limits, indicating unusual or outlier residuals. These points can provide insights into possible issues with the model, such as unmet assumptions, presence of outliers, or influence of atypical observations. Setting upper and lower limits in the temperature residual plots aids in detecting anomalous or discrepant behaviors in relation to the fitted model. This enables a more in-depth analysis and helps identify potential improvements or necessary adjustments to the regression model. Identifying these unusual behaviors can provide valuable insights into the studied system or phenomenon.

For the present study, the Tukey method was used. The method is based on calculating the interquartile range (IQR), which is the difference between the third quartile (Q3) and the first quartile (Q1) of a dataset. The first quartile (Q1) is the value that divides the data into the lowest 25% of values, and the third quartile (Q3) divides the data into the highest 25% of values. The IQR is calculated as:

$$IQR = Q3 - Q1$$

To identify outliers using the Tukey method, upper and lower limits are typically defined as multiples of the IQR above and below Q3 and Q1, respectively. A value beyond these limits is considered an outlier. The limits are calculated as follows:

$$Sup.limit = Q1 - (k * IQR)$$

$$Inf.limit = Q3 + (k * IQR)$$

Where the parameter "k" is a factor that controls the sensitivity in outlier detection. Common values for "k" are 1.5 or 3. Values beyond the upper and lower limits are considered outliers.

3. RESULTS

3.1 Model metrics

The metrics of computational analysis models have the main function of evaluating and quantifying the performance of the models in relation to the test or validation data. These metrics provide objective measures that allow for model comparison and selection and identify areas where the model may need improvement. The metrics vary depending on the type of problem (classification, regression, etc.) and the specific goal of the analysis. As for the specific analysis of linear regression models, the most common metrics for such models have been selected [12], these include R^2 (determination coefficient), MAE (mean absolute error) and RMSE (mean square error).

3.1.1 Gearbox

As can be observed in Table 6, it is noticeable that the model performed well for this major component, showing good metric values when tested on healthy turbines. This indicates that the model could accurately predict a wind turbine's ideal behavior. However, the metrics are not as good when tested on turbines with anomalies, which is exactly what is expected. Since the model is expecting values similar to those it was trained on, when anomalous values occur, the metrics will yield poor results due to the deviation of the predicted temperature from the actual temperature (increased temperature residue).

GEARBOX WITH / WITHOUT ANOMALY

PARAMETER	RNE (With / Without Anomalies)	RFO (With / Without Anomalies)	KNN (With / Without Anomalies)
R^2	0.490 / 0.95	0.308 / 0.97	0.228 / 0.92
Mean absolute error (MAE)	0.670 / 0.320	0.89 / 0.370	1.0 / 0.4
Mean square error (RMSE)	1.053 / 0.497	1.226 / 0.61	1.294 / 0.66

Table 6: Metrics for the computational model for the gearbox component with and without anomalies.

3.1.2 Main Bearing

The same analysis that was done for the gearbox component can also be done for the main bearing. Therefore, good results were obtained for this major component, as can be seen in Table 7 below.

<table>
<tr><th colspan="4">ROLAMENTO PRINCIPAL COM / SEM ANOMALIA</th></tr>
<tr><th>PARAMETER</th><th>RNE (With / Without Anomalies)</th><th>RFO (With / Without Anomalies)</th><th>KNN (With / Without Anomalies)</th></tr>
<tr><td>R^2</td><td>0.703 / 0.999</td><td>0.736 / 0.913</td><td>0.655 / 0.920</td></tr>
<tr><td>Mean absolute error (MAE)</td><td>3.29 / 2.425</td><td>3.41 / 2.62</td><td>3.2 / 2.6</td></tr>
<tr><td>Mean square error (RMSE)</td><td>10.054 / 4.193</td><td>10.153 / 4.294</td><td>9.911 / 4.278</td></tr>
</table>

Table 7: Metrics for the computational model for the main bearing component with and without anomalies.

3.1.3 Generator

The same analysis that was done for the gearbox component and the main bearing can also be done for the generator component. Therefore, good results were obtained for this major component, as can be seen in Table 8 below.

<table>
<tr><th colspan="4">GERADOR COM / SEM ANOMALIA</th></tr>
<tr><th>PARAMETER</th><th>RNE (With / Without Anomalies)</th><th>RFO (With / Without Anomalies)</th><th>KNN (With / Without Anomalies)</th></tr>
<tr><td>R^2</td><td>0.488 / 0.77</td><td>0.468 / 0.722</td><td>0.449 / 0.67</td></tr>
<tr><td>Mean absolute error (MAE)</td><td>2.00 / 1.916</td><td>2.21 / 2.24</td><td>2.4 / 2.4</td></tr>
<tr><td>Mean square error (RMSE)</td><td>7.023 / 4.603</td><td>7.156 / 5.05</td><td>7.281 / 5.505</td></tr>
</table>

Table 8: Metrics for the computational model for the generator component with and without anomalies.

3.2 Residual temperatures time series plot

To demonstrate the results, three wind turbines representing each major component studied were selected to show the behavior of the temperature residual over time and the trend of the residual curve until there is an actual intervention by the wind turbine maintenance team. Firstly, the wind turbine representing the gearbox component with anomalies will be presented, followed by the wind turbine representing the main bearing, and finally, the wind turbine showing the generator's behavior.

In the plots of Figures 15, 16, and 17, it is possible to observe the increasing trend of residuals as they approach the failure point, which is favorable for the analyst as it shows that the model is well-trained. Although the visualizations for each component are different, they exhibit the same behavior of increasing residual as the failure period approaches. In Figure 15, there is a slight growth in the residual, not as clearly evident as in Figures 16 and 17 in Figure 16, there is a clear visualization of the residual growth as it approaches the failure, but there is a strong presence of false positives at the beginning of the series. These figures are preliminary approaches to the study and their main objective is to observe if the computational model identified a change in the residual behavior as the anomaly approached for the large component.

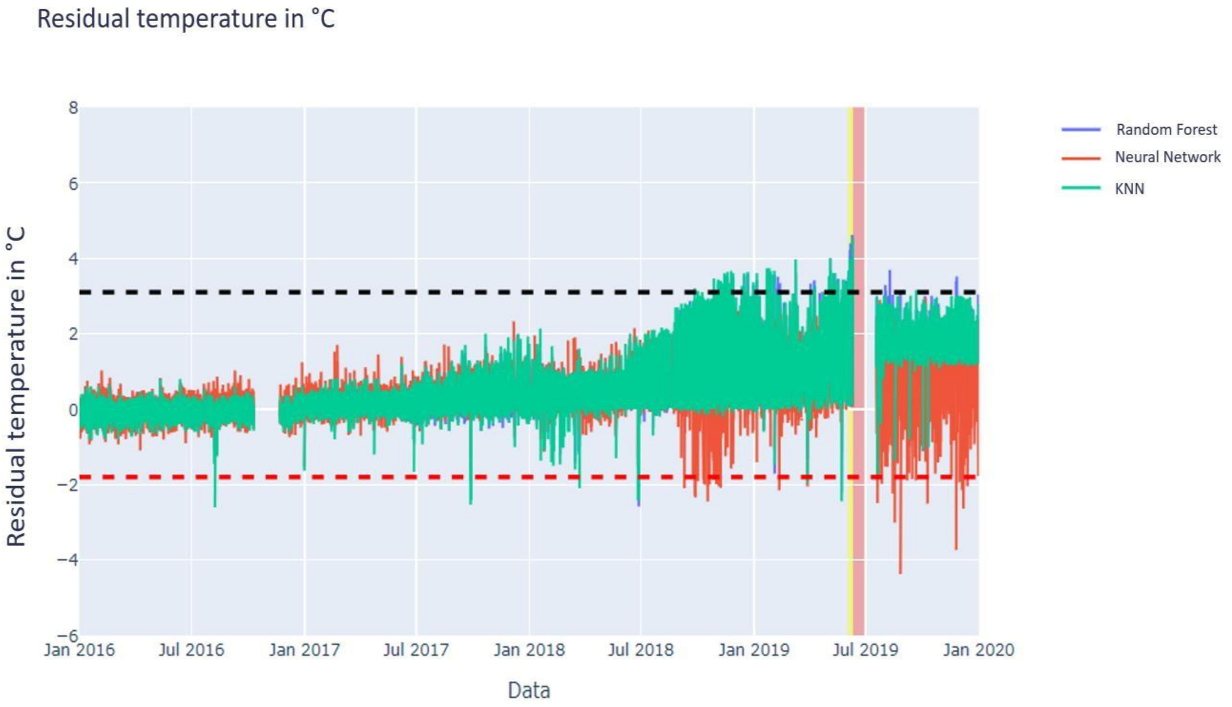


Figure 15: Residual Plot for Wind Turbine 1.

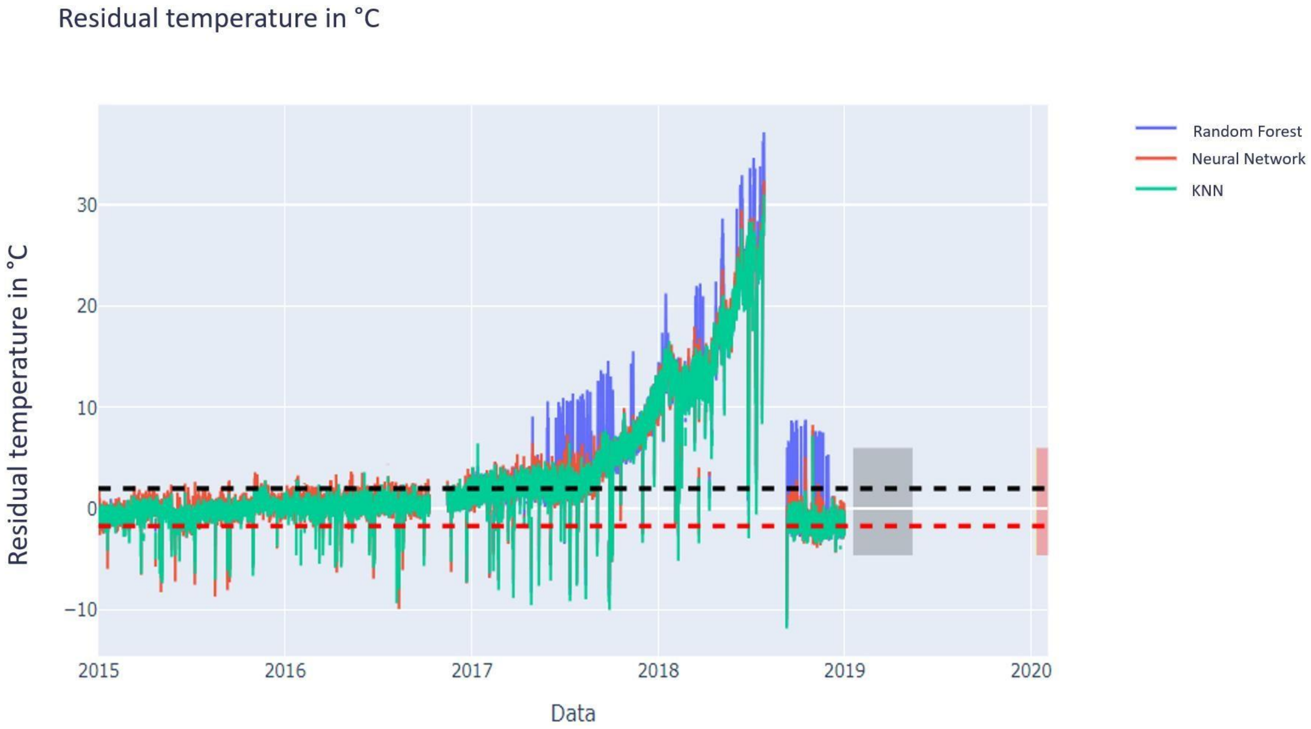


Figure 16: Residual Plot for Wind Turbine 2.

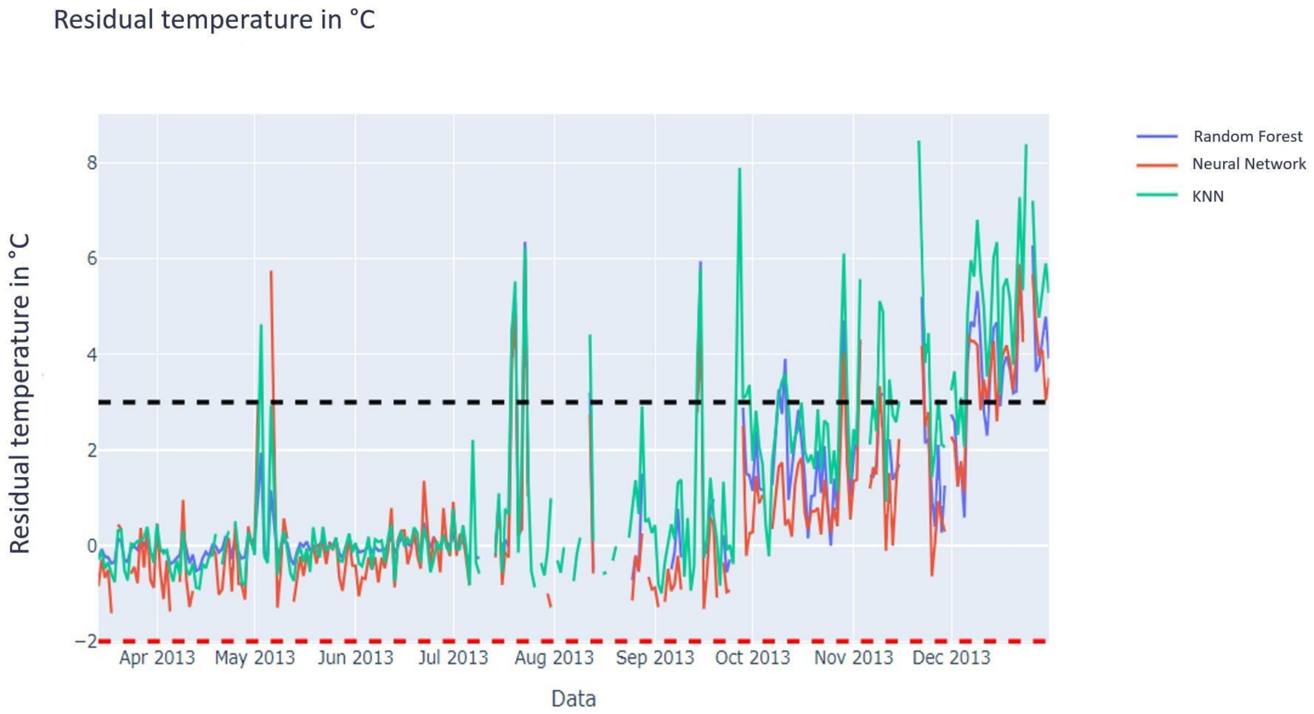
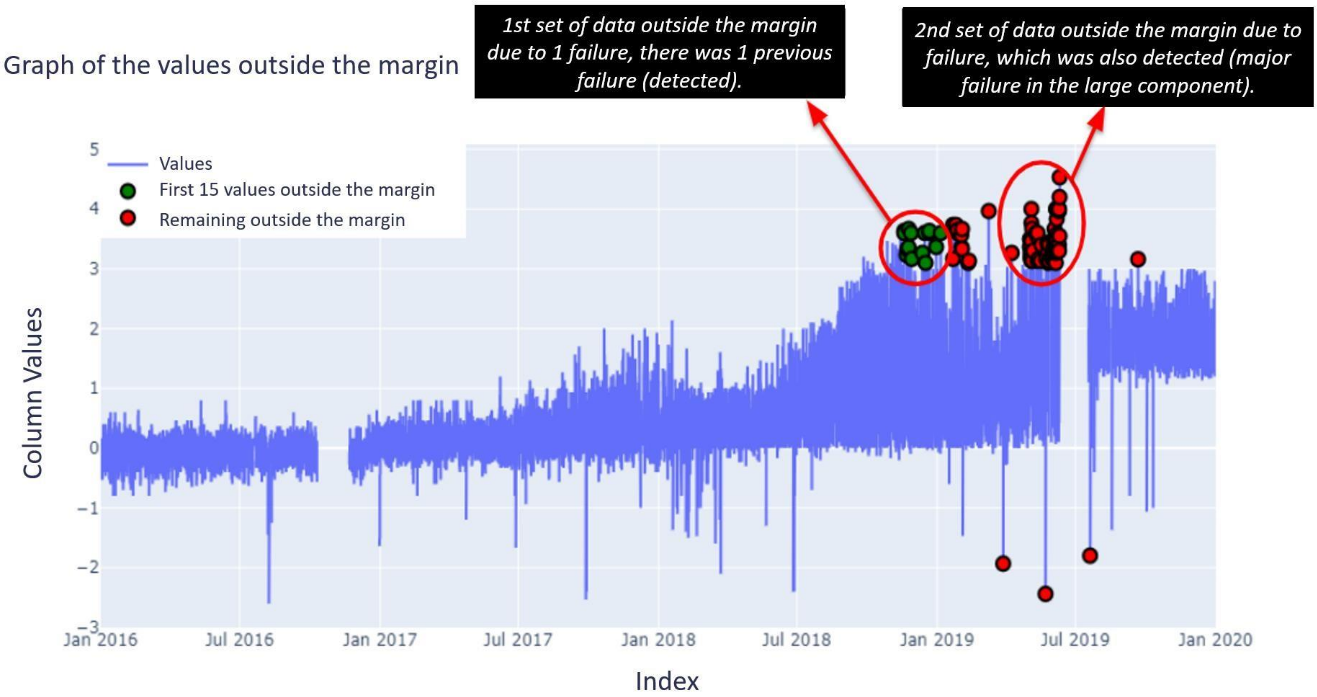


Figure 17: Residual Plot for Wind Turbine 3

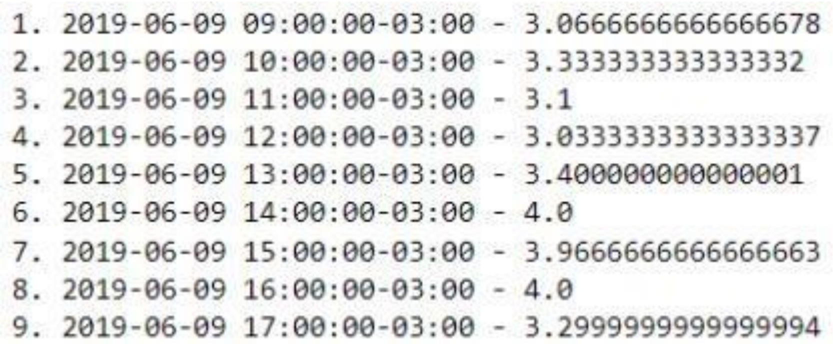
In addition to observing the model's performance through the residual plot, it also provides the sequence of data points that were outside the thresholds, indicating the date and time when the values remained outside the limits. This information is beneficial as it estimates a growth curve based on the date when the residuals started to exceed the upper or lower determined limits. Figures 18 and 19 highlight the sequences outside the thresholds and provide examples of how the model automatically returns the sequence that is outside the established limits.



1st set of data outside the margin due to 1 failure, there was 1 previous failure (detected).

2nd set of data outside the margin due to failure, which was also detected (major failure in the large component).

Figure 18: 1st Wind Turbine exemplifying the visualization of sequences outside the thresholds. Source: Author.



```
1. 2019-06-09 09:00:00-03:00 - 3.0666666666666678
2. 2019-06-09 10:00:00-03:00 - 3.333333333333332
3. 2019-06-09 11:00:00-03:00 - 3.1
4. 2019-06-09 12:00:00-03:00 - 3.0333333333333337
5. 2019-06-09 13:00:00-03:00 - 3.400000000000001
6. 2019-06-09 14:00:00-03:00 - 4.0
7. 2019-06-09 15:00:00-03:00 - 3.9666666666666663
8. 2019-06-09 16:00:00-03:00 - 4.0
9. 2019-06-09 17:00:00-03:00 - 3.2999999999999994
```

Figure 19: Sequence of data outside the thresholds (date, time, and temperature residue value).

4. CONCLUSIONS

This paper focuses on predictive maintenance, aiming to reduce unexpected expenses from unforeseen failures and preventive maintenance in wind turbines.

In the first section, the scenario of wind energy generation in Brazil was presented, along with the definition and applications of machine learning, which is the AI model being used in this study. It also builds upon previous studies and introduces the data collection and information system (SCADA system).The results presented demonstrate that with the aid of the SCADA system, it is possible to improve the current scenario of wind energy generation in Brazil and worldwide. The cost of wind turbines and their maintenance are the major challenges faced today, so this study can significantly reduce these costs.

In conclusion, the study presented a comprehensive approach to detecting failures in wind turbines using temperature analysis and statistical models. By monitoring the temperature of the main components, deviations from normal operating conditions can be detected, enabling proactive and predictive maintenance, and minimizing downtime. The integration of machine learning techniques such as neural networks, random forests, and k-nearest neighbors enhances the accuracy of failure detection. The correlation with the current state of wind energy generation highlights the importance of effective failure detection techniques in ensuring the reliability and efficiency of wind turbines. The use of the Python programming language, along with relevant libraries for data analysis and visualization, facilitates the implementation of the proposed methodology. Future research can focus on refining the statistical models and exploring additional data sources for comprehensive failure detection systems in wind turbines. The models' performance and the residuals' behavior indicate abnormal patterns during failures, while remaining linear and within the predetermined thresholds during healthy periods for the selected wind turbines. The residual plots demonstrate that for each major component, there is a time interval in which there is a variation in the temperature variables. The computational models (neural network and random forest) are able to predict failures with a lead time of 80 to 100 days for the gearbox, 90 to 120 days for the main bearing, and 60 to 80 days for the generator.

Regardless of the specific major component studied, the machine learning model shows variations in the results for the selected wind turbines and input and target variables. However, the model is designed to detect anomalies in the major components (gearbox, main bearing, and generator) using three machine learning techniques and input variables most relevant to the specific as shown in Tables 6, 7, and 8, indicate abnormal patterns during periods of failures and linear behavior within the predetermined thresholds during healthy periods.

5. ACKNOWLEDGEMENTS

The information in this study was obtained during the development of a Research and Development - R&D project supported by Rio Amazonas SA and approved by the Brazilian Regulatory Agency ANEEL. The second author thanks CNPq for the support through the post-doc grant n. 200820/2022-2 under Call 26/2021. The third author thanks CNPq for the support through the productivity grant n. 303417/2022-6.

6. REFERENCES

01. González, M. O. A., Gonçalves, J. S., & Vasconcelos, R. M. (2017, January 20). Sustainable development: Case study in the implementation of renewable energy in Brazil. ELSEVIER.

02. Silva, N. F. da, Rosa, L. P., Freitas, M. A. V., & Pereira, M. G. (2013, April 02). Wind energy in Brazil: From the power sector's expansion crisis model to the favorable environment. ELSEVIER.

03. Murgia, A; Verbeke, R; Tsiporkova, E; Terzi, L.; Astolfi, D. Discussion on the Suitability of SCADA-Based Condition Monitoring for Wind Turbine Fault Diagnosis through Temperature Data Analysis. Sensors 2021, 21,5867.

04. Turnbull, A.; Carrol, J.; McDonal, A. A Comparative Analysis of the Variability of Temperature Thresholds through

Time for Wind Turbine Generators Using Normal Behaviour Modelling. Energies 2021, 14, 3211.

05. Zaher, A.; McArthur, S.; Infield, D.; Patel, Y. Online wind turbine fault detection through automated SCADA data analysis. Wind Energy: Int. J. Prog. Appl. Wind Power Convers. Technol. 2009, 12, 574–593.
06. Turnbull, A.; Carroll, J.; McDonald, A. A comparative analysis on the variability of temperature thresholds through time for wind turbine generators using normal behaviour modelling. Energies 2022, 15, 5298
07. Reimann, C., Filzmoser, P., & Garret, R. G. (2005, February 4). Background and threshold: critical comparison of methods of determination. ELSEVIER.
08. Wang, C., Viswanathan, K., Choudur, L., Talwar, V., Satterfield, W., & Schwan, K. (2011, August 18). Statistical techniques for online anomaly detection in data centers. IEEE.
09. Han, B., Xie, H., Shan, Y., Liu, R., & Cao, S. (2021, November 19). Characteristic Curve Fitting Method of Wind Speed and Wind Turbine Output Based on Abnormal Data Cleaning. ATAMI.
10. Loca, A. L. S. A methodology for experimental evaluation of machine learning approaches for fault diagnosis based on vibration signals. Master's defense, Federal University of Espírito Santo, Vitória, Espírito Santo, Brazil, 2020.
11. Encalada-Dávila, Á.; Puruncajas, B.; Tutivén, C.; Vidal, Y. Wind turbine main bearing fault prognosis based solely on scada data. Sensors 2021, 21, 2228.
12. Magalhães, N. A. F. (2011). Wind Farm Performance Analysis Support System (Master’s thesis). Faculdade de Engenharia da Universidade do Porto, March 2011, p. 17.
13. Terra, N. (2023, April 4). Wind energy in Brazil breaks records and creates jobs. Retrieved from [https://www.airswift.com/blog/wind-energy-brazil]
14. Batista, B. C. F. (2012). Solutions of Differential Equations Using Multilayer Neural Networks with the Steepest Descent and Levenberg-Marquardt Methods (Master’s thesis). UFPA.
15. Vinicius. (2017). Artificial Neural Networks. Monolito Nimbus. Retrieved from [https://www.monolitonimbus.com.br/redes-neurais-artificiais/]
16. Bhavsar, D. (2020, May 6). Dispelling Myths: Deep Learning vs. Machine Learning. Merkle Blog. Retrieved from [https://www.merkle.com/blog/dispelling-myths-deep-learning-vs-machine-learning]
17. Donges, N. (2023, March 14). Random Forest: A Complete Guide for Machine Learning. Builtin.com. Retrieved from [https://builtin.com/data-science/random-forest-algorithm]
18. Msigwa, O. J. (2022, November 15). Data Science and Machine Learning (Part 09): The K-Nearest Neighbors Algorithm (KNN). MQL5. Retrieved from [https://www.mql5.com/en/papers/11678]

7. RESPONSIBILITY NOTICE

The author(s) is (are) the only responsible for the printed material included in this paper.