Report for the Machine Learning Project

Jakub Wojciechowski, Tymoteusz Tomczak
Poznań Institute of Technology
Introduction to Artificial Intelligence - final project

28.01.2025

1 Discussion of Fault Detection Methods in Industrial Motors

Discussion of fault detection methods in industrial motors.

1.1 First Method: Analysis Using Linear Regression and Random Forest [1]

The first method of temperature data analysis involves using linear regression as a baseline model and the Random Forest algorithm, which better captures non-linear relationships in the data, surpassing linear regression in accuracy.

Strengths:

- Random Forest effectively models non-linearities and works efficiently even with a larger number of features.
- Interpretability of results through feature importance evaluation.

Weaknesses:

- Higher computational demands compared to simpler models, such as linear regression.
- Random Forest may overfit on small or imbalanced datasets.

1.2 Second Method: K-NN and SVM Classifiers[2]

In the work "A Machine Learning Framework for Bearing Fault Detection in Three-Phase Induction Motors," K-Nearest Neighbors (K-NN) and Support Vector Machine (SVM) classifiers were proposed to predict temperature in PMSM motors. The effectiveness of both algorithms was compared, indicating:

[1] S. Sawant, “Electrical Motor Temperature: PMSM,” 2021. Available online: kaggle.com.

[2] W. Rohouma, A. Zaitouny, M. F. Wahid, H. Ali and S. S. Refaat, “A Machine Learning Framework for Bearing Fault Detection in Three-Phase Induction Motors,” 2024 4th International Conference on Smart Grid and Renewable Energy (SGRE), Doha, Qatar, 2024, pp. 1-5, doi: 10.1109/SGRE59715.2024.10429024.

- K-NN performs well with local data dependencies.
- SVM shows better generalization ability for more complex datasets.

Strengths:

- **K-NN**: Simple to implement, requires no extensive training process, and performs well with local patterns in the data.
- **SVM**: High prediction accuracy, especially for complex problems, due to the use of non-linear kernel functions.

Weaknesses:

- **K-NN**: High computational complexity during prediction, particularly for large datasets.
- **SVM**: Requires careful parameter tuning (e.g., kernel function) and is less efficient with very large datasets.

1.3 Third Method: Thermal Imaging Using BNN [3]

In the work "Diagnosis of the three-phase induction motor using thermal imaging," the authors proposed a custom solution using Bayesian Neural Networks (BNN), a nearest neighbors classifier, and a centroid algorithm.

BNN Features:
Strengths:

- Uncertainty modeling: Probabilistic approaches allow estimating result confidence.
- Flexibility: Ability to model complex data dependencies.
- Robustness: Less prone to overfitting compared to traditional neural networks.
- Interpretability: Probabilistic weight distributions provide additional insights into feature importance.

Weaknesses:

- High computational complexity (e.g., Monte Carlo techniques are required).
- High data demands to accurately estimate weight distributions.
- Difficulties in scaling to very large datasets.

[3] Adam Glowacz, Zygfryd Glowacz, Diagnosis of the three-phase induction motor using thermal imaging, Infrared Physics Technology, Volume 81, 2017, Pages 7-16, ISSN 1350-4495, 10.1016/j.infrared.2016.12.003, www.sciencedirect.com

Centroid Algorithm Features:

Strengths:

- Fast operation and ease of implementation.
- Efficiency for well-separated data.

Weaknesses:

- Low accuracy for complex datasets.
- Lack of adaptability and sensitivity to data distribution.

2 Context and Program Assumptions

The program aimed to test three methods and determine which algorithm performs best under various conditions. The goal was to assess whether the model, based on process variables, could predict motor overheating, which might lead to damage.

2.1 Dataset

The data used in the project comes from Kaggle, collected by scientists at Paderborn University. It was reduced to 17,000 records and included 12 columns:

1. ID
2. U_q – q-axis voltage component
3. Coolant – coolant temperature
4. Stator_winding – stator winding temperature
5. U_d – d-axis voltage component
6. Stator_tooth – motor tooth temperature
7. Motor_speed – motor rotational speed
8. I_d – d-axis current component
9. I_q – q-axis current component
10. Pm – permanent magnet temperature
11. Stator_yoke – yoke temperature
12. Torque – torque
13. Profile_id – measurement session ID

For our problem, the models receive all fields except temperature fields and Profile_id (as it does not affect measurement results since we use only one ID for learning). Models predict magnet temperature by default. For logistic regression, the data was standardized using `StandardScaler()`, while SVM data was normalized with `MinMaxScaler()`.
With some modifications, the program can also predict other process variables by changing ignored columns and the target variable (y). At the end of the program, decision tree graphs, accuracy metrics for each method, confusion matrices, and correlation matrices are displayed.

3 Presentation of Results

3.1 Accuracy Results:

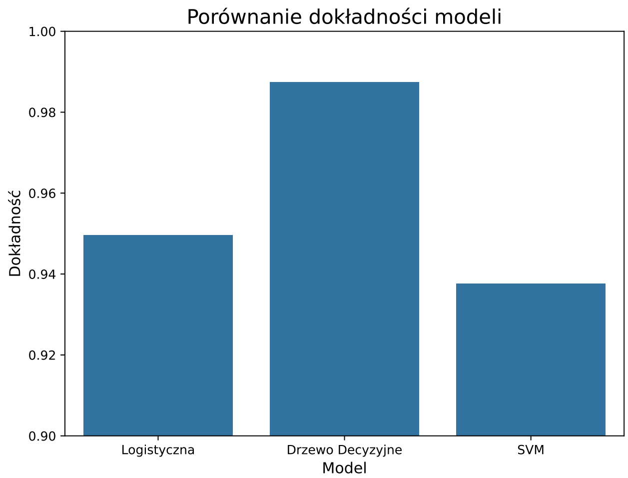

Figure 1: Model accuracy graphs

- **Decision Tree**: Accuracy 98.7%.
- **Logistic Regression**: Accuracy 94.9%.
- **SVM**: Accuracy 93.8%.

3.2 Decision Tree

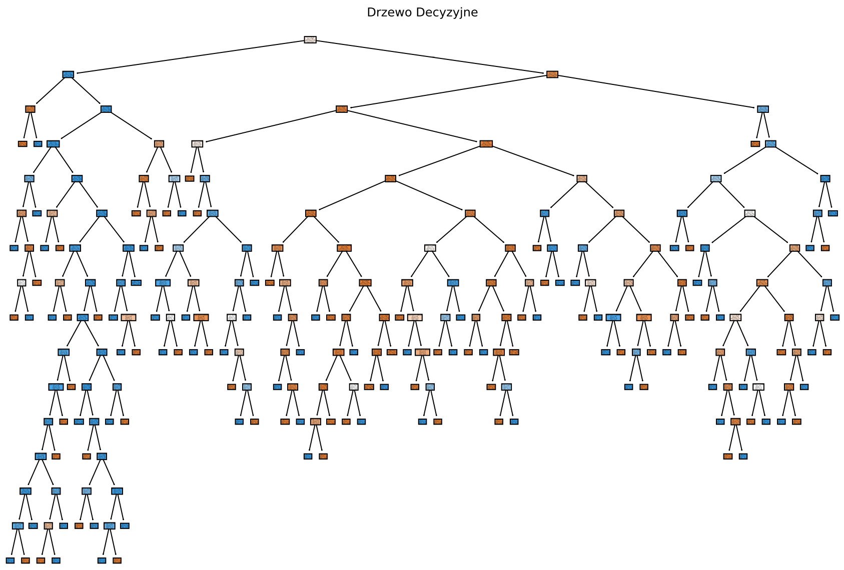

Figure 2: Decision tree

3.3 Confusion Matrices

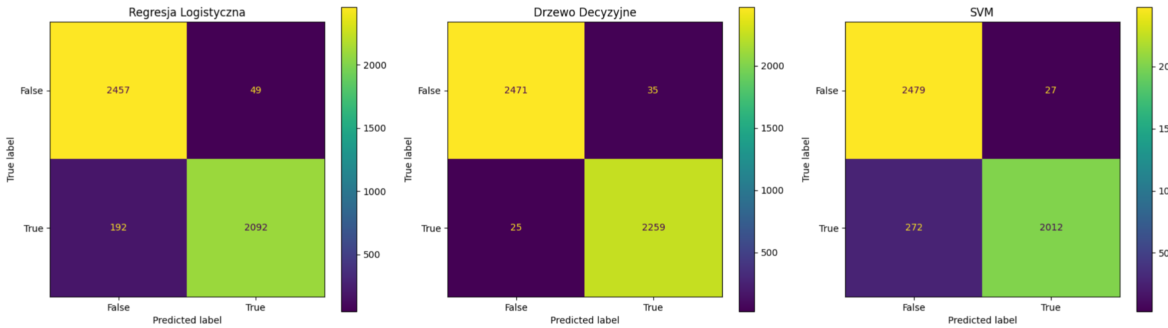

Figure 3: Confusion matrices

3.4 Correlation

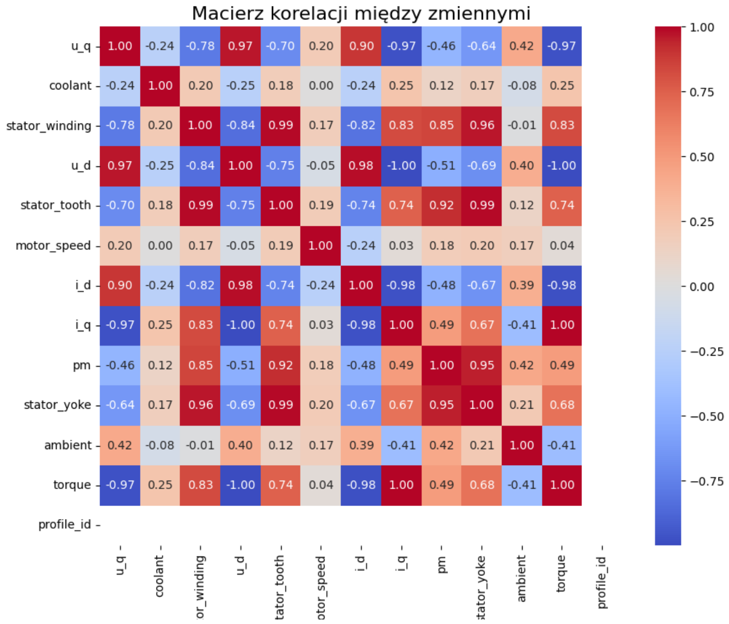

Figure 4: Correlation matrix

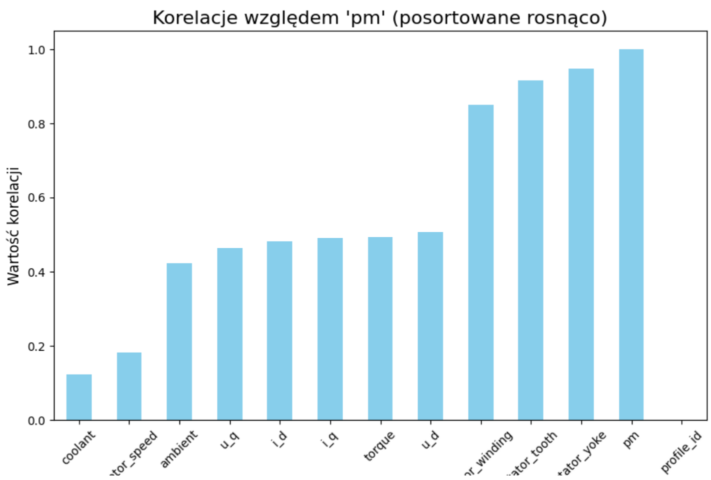

Figure 5: Sorted correlation representation

4 Conclusions

All models achieved very satisfactory levels of accuracy. Despite the highest accuracy, the decision tree does not seem adequate for such problems as it becomes very extensive and difficult to analyze. It is also not possible to definitively determine which of the remaining models performs better, as our test was based on fewer samples than would be the case in an industrial setting, and both models exhibit very similar predictive accuracy. Correlation matrices reveal that all temperatures are directly proportional to each other, with very high coefficients near unity. Thus, removing any temperature measurement increases the problem's complexity for the models. The least useful measurement appears to be rotational speed, as its correlation with any variable does not exceed 0.2. In the context of industrial application, it would be necessary to provide direct measurements of voltage, current, and torque. If it were possible to measure any temperature parameter, model accuracy would increase significantly due to correlation. Such solutions should be considered for new designs, where data acquisition is a fundamental assumption, enabling the prediction of potential faults. For example, instead of predicting permanent magnet temperature, predictions could focus on components subject to friction, such as bearings, which, when overheated, cause malfunctions and financial losses in the industry.