Bayesian Reasoning and Machine Learning
Assignment 1

by

Michel Mariën (852674168)

Course code: IM0892

Open Universiteit
de beste! www.ou.nl

Contents

Acronyms

CPT Conditional Probability Table. 3, 5

1. Introduction

1.1. Motivation

As a maintenance engineer, my job is keeping uptime as high as possible by on understanding and preventing machine breakdowns. Unfortunately, not all failures are easy to predict or troubleshoot. Therefore I want to see if it is possible to create a Bayesian Network-based tool for aiding in troubleshooting machine breakdowns.

1.2. Data

The data in this assignment is based on the *AI4I 2020 dataset*. The dataset is a synthetic dataset that reflects real machine sensor data encountered in industry and is recommended for testing machine learning methods such as classification, regression, causal-discovery [Rep20]. The dataset consists of 14 columns and 10.000 rows (see table 1).

Variable	Role	Type	Missing	Description
UID	ID	Integer	no	Unique identifier
Product ID	ID	Cat.	no	Product variants + serial number
Type	Feature	Cat.	no	Letters L, M, or H for product variants
Air temperature	Feature	Cont.	no	Air temperature in degrees K
Process temperature	Feature	Cont.	no	Process temperature in degrees K
Rotational speed	Feature	Integer	no	Calculated revolutions per minute.
Torque	Feature	Cont.	no	Torque values in Nm
Tool wear	Feature	Integer	no	tool wear in minutes
Machine failure	Target	Integer	no	Indication of machine failure
TWF	Target	Integer	no	Failure due to Tool wear
HDF	Target	Integer	no	Failure due to heat dissipation
PWF	Target	Integer	no	Failure due to power failure
OSF	Target	Integer	no	Failure due to overstrain failure
RNF	Target	Integer	no	Failure due to random failure

Table 1: Dataset description, for a more extensive description, see [Rep20]

1.3. Data preprocessing

Before the data is suitable for this assignment, the data has to be preprocessed. For example, all continuous variables (’Tool wear’, ’Air temperature’, ’Torque’, ’Process temperature’, ’Rotational speed’) will have to binned to discrete values and all target columns will have to be converted to Boolean before constructing the Bayesian Network. For an overview of all preprocessing, see [Mar25].

The dataset consists of 10.000 rows with only 339 machine failures, which seems reasonable since the different rows represents measurement of all variables based on time and not just machine failures. However, this means there is a risk on very low probability values in the Conditional Probability Table (CPT). Therefore, after converting the variables (but before binning), the dataset will be examined to see if the amount of data can be reduced without changing the prior probabilities too much. For example, wear-induced failures are not expected early in a system lifespan because wear-induced failures take time to develop. Thus this variable will investigated to see if the data with little wear can be removed. The variable is plotted in fig. 1.

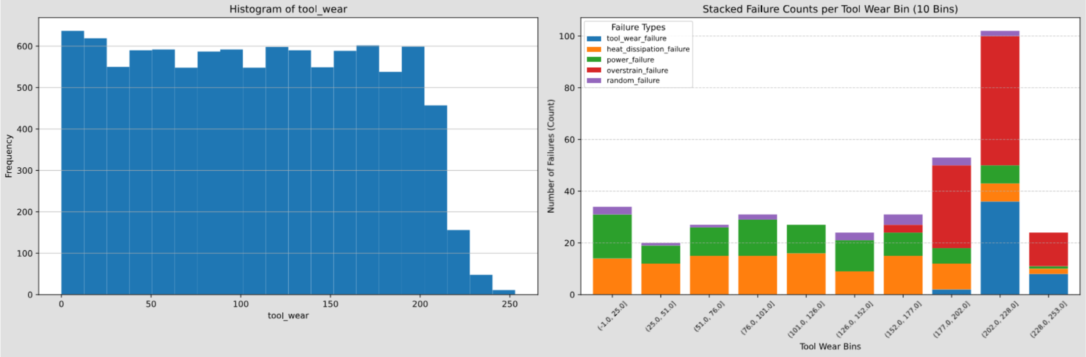

Figure 1: Histogram and stacked bargraph for *Tool Wear*

In fig 3, it can be observed that *Tool wear* only has a direct influence on target variables *Overstrain failure* and *Tool wear failure*. Looking at variable in fig. 1, it can be observed that no *Tool wear* failures occur before the bin (152-177] and only three *Overstrain failure* (checking the data reveals that the first *Tool wear failure* occurs at 198 min and the first *Overstrain failure* at 172 min). Eliminating the data in this 'wear-induced-failure-free'-period from the dataset should thus have little influence on probability on 'wear-induced-failures. However, in that same time period, several other failure modes are present. Removing the data before 172 would also remove those failure modes and reduce the total number of failures from a total of 339 to 164. Since this will leave out to much machine breakdowns, all data will be retained for modelling and create the Bayesian Network and Inference in the next sections.

2. Model

The basic layout of the network is described in [Rep20]. Herein, the authors specify the direct relations between the different feature-variables and the target-variables. They also describe the specific interactions and thresholds between the variables and how these lead to machine failure. However, the thresholds will be ignored in this report and only the direct variable-relations will be used. The model is given in the figures 2 and 3.

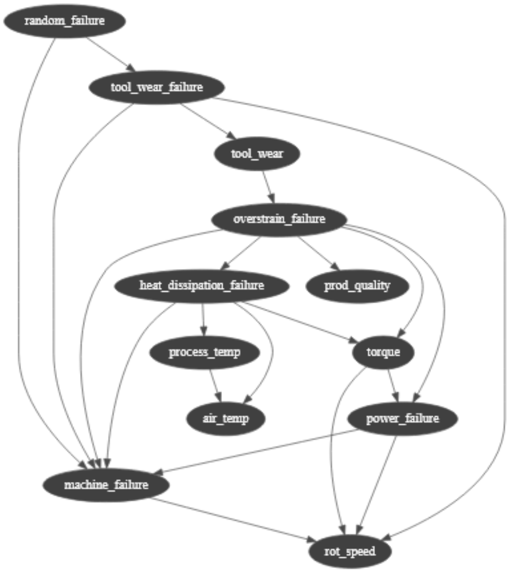

Figure 2: Network learned from the data

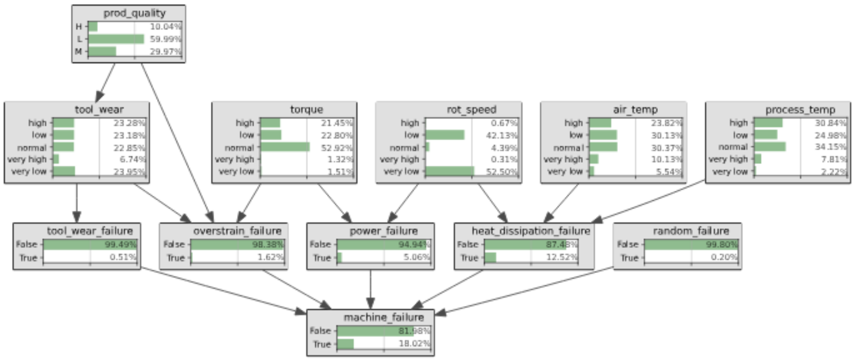

Figure 3: Modified network after manually adding/blocking arcs, supplemented with prior probabilities

The model in 3 is created by letting the model learn the general structure from the data and modifying the model to align with [Rep20] by adding mandatory arcs and block any non-allowable arcs using the PyAgrum functions *.addMandatoryArc()* and *.addForbiddenArc()* [pT25]. After generating the network, the Conditional Probability Table (CPT)'s can be visualized. However, since the feature-variables are all binned to 5 discrete values, the tables have become very long. Therefore, the tables with the feature variables are given in fig. 4 and for the Conditional Probability Table (CPT)'s of the target variables, the reader is referred to the notebook [Mar25].

prod_quality

prod_quality		
H	L	M
0.1004	0.5999	0.2997

torque

torque				
high	low	normal	very high	very low
0.2145	0.2280	0.5292	0.0132	0.0151

air_temp

air_temp				
high	low	normal	very high	very low
0.2382	0.3013	0.3037	0.1013	0.0554

tool_wear

	tool_wear				
prod_quality	high	low	normal	very high	very low
H	0.2282	0.2391	0.2302	0.0655	0.2371
L	0.2335	0.2276	0.2293	0.0701	0.2395
M	0.2328	0.2378	0.2262	0.0626	0.2405

rot_speed

rot_speed				
high	low	normal	very high	very low
0.0067	0.4213	0.0439	0.0031	0.5250

process_temp

process_temp				
high	low	normal	very high	very low
0.3084	0.2498	0.3415	0.0781	0.0222

Figure 4: Conditional Probability Table (CPT)'s for the feature variables

3. Inference

A maintenance engineer or any troubleshooting agent in general wil try to determine the root cause of a breakdown as soon as possible. This can be done by examining sensor readings prior to failure. By systematically checking the measurements, costly disassembly and possibly replacing the wrong parts might be avoided. To this aim, the network could be used as follows:

3.1. Scenario 1: High stress

A machine has failed. During troubleshooting the engineer observed that the machine had high rotational speed and torque prior to failure.

Potential question for the model:
"Given high rotational speed, high torque and machine failure, what is the probability of 'overstrain_failure'"

- Rotational speed *(rot_speed)* is high
- Torque *(torque)* is high
- Machine failure *(machine_failure)* is true

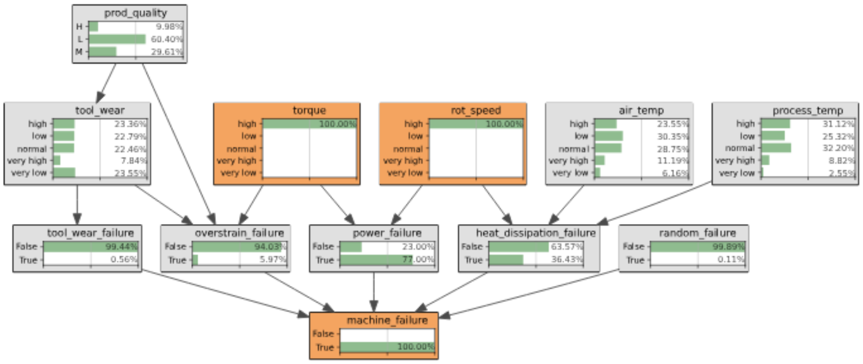

Figure 5: Posterior probabilities in the **High stress scenario**

In fig. 5, the probability on *'overstrain_failure'* has risen from 1,62% (prior probability) to 6,38% (posterior probability)

3.2. Scenario 2: High temperatures

A machine has failed. During troubleshooting the engineer observed that both air and process temperatures were very high prior to failure.

Potential question for the model:
"Given both air and process temperatures were 'very high' during machine failure, what is the likelihood of 'heat_dissipation_failure'"

- Air temperature *(air_temp)* is very high
- Process temperature *(process_temp)* is very high
- Machine failure *(machine_failure)* is true

In fig. 6, the probability on *'heat_dissipation_failure'* has dropped from 12,52% (prior probability) to 9,14% (posterior probability). This means that the chance on this failure mode has dropped, despite the high temperatures.

3.3. Scenario 3: High temperatures revisited

During the previous investigation, the engineer realized he made a mistake, the process temperature was actually low. He also noticed that the rotational speed was low as well.

Potential question for the model:
"Given the additional information, what is the likelihood of 'heat_dissipation_failure'"

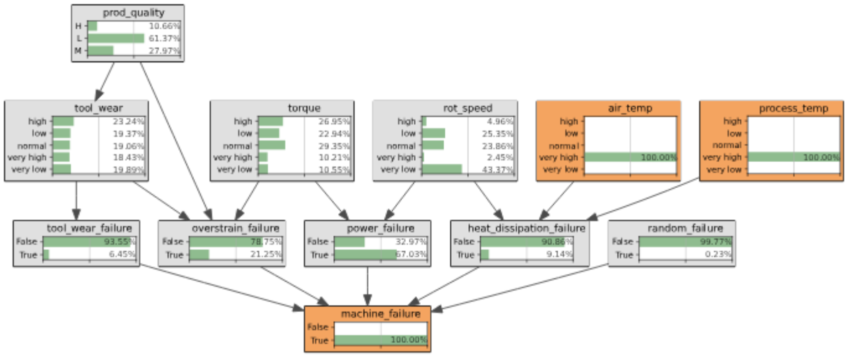

Figure 6: Posterior probabilities in the **High temperatures scenario**

- Air temperature *(air_temp)* is very high
- Process temperature *(process_temp)* is low
- Rotational speed *(rot_speed)* is low
- Machine failure *(machine_failure)* is true

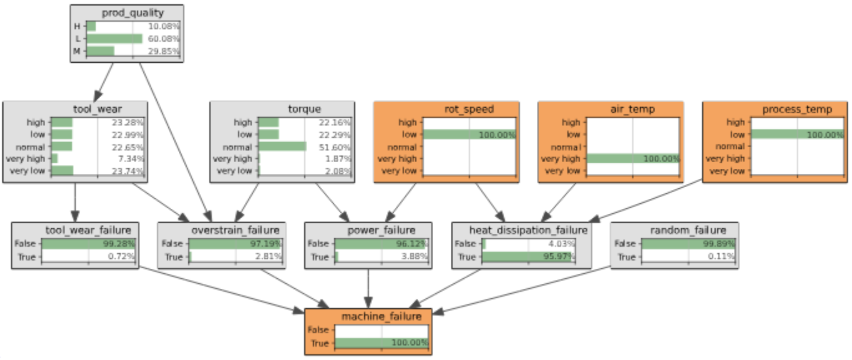

Figure 7: Posterior probabilities in the **High temperatures scenario**

In fig. 7, the probability on *'heat_dissipation_failure'* has increased from 12,52% (prior probability) to 95,97% (posterior probability). This means that the chance of this failure mode has dramatically increased. This also confirms the statement that *'heat_dissipation_failure'* is caused by a temperature difference between air temperature/process temperature and a low rotational speed [Rep20].

4. Discussion

This report describes the creation of Bayesian network to aid in troubleshooting. Although the network is very basic, it confirms statements how certain variables interact and cause a machine breakdown. Such a model could also be used in preventing breakdowns by putting alarms on individual sensor measurements and feeding it live to a model like the one described in this report. This can help the operator or mechanic in preventing the failure by taking the correct actions in time. It is advisable to collect additional measurement variables before making such an attempt.

From a technical perspective, an interesting further development for the model could be to include 'noisy' gates to reduce the number of features that have to be calculated. For example, the inference-example around *'heat_dissipation_failure'* could perhaps also be modeled as a gate with the specific combination of the variable values as determining the truth value of the gate.

References

[Mar25] Michel Marien. Source code for brml assignment 1, 2025. Accessed: 2025-05-19. 3, 5

[pT25] pyAgrum Team. Introduction to pyagrum, 2025. Accessed: 2025-05-19. 5

[Rep20] UC Irvine Machine Learning Repository. Ai4i 2020 predictive maintenance dataset, 2020. 2, 4, 5, 7