ASSIGNMENT 2: Making Choices in Data

Olarte, Daniel M. December 20, 2023

BSIT-MI211 Analytic 3

1. **What variables might be relevant to the decision?**

   Dependent Variable:
   - OEE (Overall Equipment Effectiveness)
   - On-time performance
   - Cost of maintenance per flight hour

   Independent Variables:
   - Sensor data from aircraft components
   - Historical maintenance records
   - Mechanical data
   - Manufacturer data on component lifetimes and failure rates

   1.1 What are the hypotheses about how to improve the data?

   Hypothesis 1: Improved Data Integration
   - Integrating data from multiple sources improves predictive capabilities.

   Hypothesis 2: Timeliness of Data
   - Real-time flight data will contribute to more accurate predictions.

   1.2 What variables are needed to test the hypotheses?

   - Binary Variables and Continuous variables

2. **What data sources can inform those variables?**

   - Real time flight data
   - Past maintenance activities details
   - Failure Rates and Component lifetimes

   2.1 What are the possible sources of the necessary information?

   - I found some credible sources on Kaggle like:
   https://www.kaggle.com/datasets/shivamb/machine-predictive-maintenance-classification
   https://www.kaggle.com/code/sharanharsoor/aircraft-predictive-maintenance

2.2 Are there any considerations that constrain which data sources can be used (timeliness, access, privacy, cost, standards, etc.)?

- Yes, some of my needed data sources are needed to buy. For now, I am still finding credible sources for this project.

3.1 How will you transform the raw data into variables suitable for modeling?

- I think there are only minimal changes that can be made to the dataset like Cleaning and Normalization. I also know that there are some more credible sources that I can get but not for now, because there are credible datasets that can be bought.

3.1 What does profiling of each data source reveal about the quality of the available data?

- Profiling of each data source can reveal:

- Number of Records
- Fields in a Record
- Data Format, Units, Range of Values
- Summary Statistics

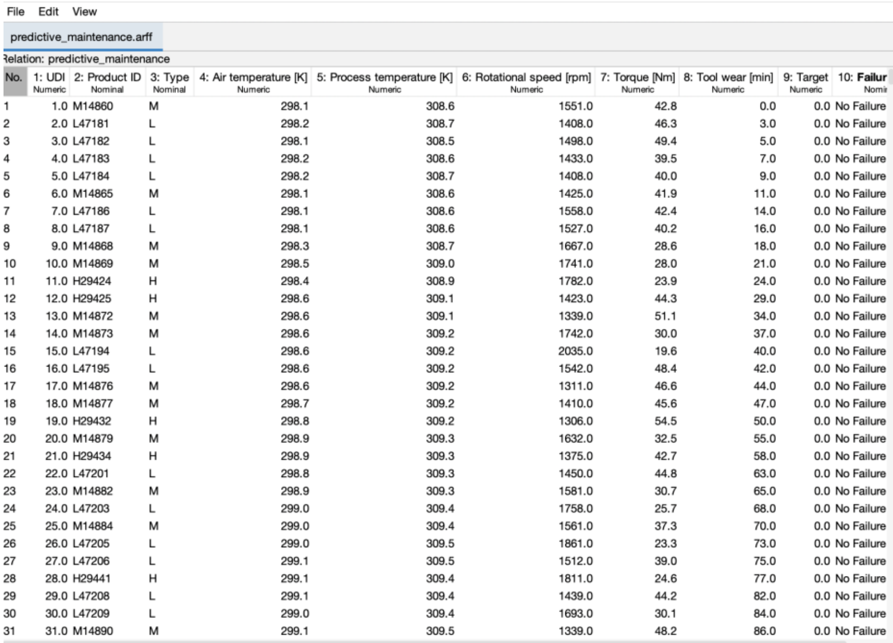
3.2 How should data from different sources be integrated? Is there a need to integrate?

- From what I found, there are two sources I can integrate to each other to make a larger dataset. By integrating it, I would create keys that would match the variables and do some consistency checks.

3.3 What additional preparation steps are necessary on the integrated data to yield variables for analytic modeling?

- I think what is nice to do is identifying outliers and creating new variables to make the dataset organized.

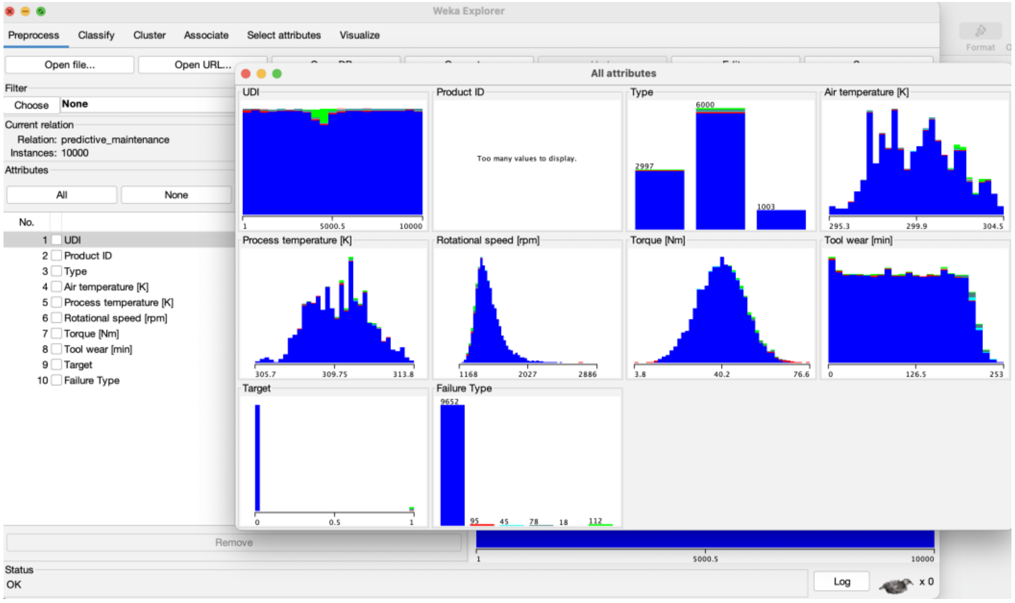
You will need to profile the data sources to complete this assignment. Profiling should include at a minimum the number of records in a data source (for each table, as applicable), the fields in a record, and for each field, the format of the data, units, the range of values, and summary statistics for numeric data. You may want to create some data visualizations to understand the range of values and spot outliers.


File
Edit
View
predictive_maintenance.arff

Relation: predictive_maintenance

No.	1: UDI	2: Product ID	3: Type	4: Air temperature [K]	5: Process temperature [K]	6: Rotational speed [rpm]	7: Torque [Nm]	8: Tool wear [min]	9: Target	10: Failur
	Numeric	Nominal	Nominal	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Nomi
1	1.0	M14860	M	298.1	308.6	1551.0	42.8	0.0	0.0	No Failure
2	2.0	L47181	L	298.2	308.7	1408.0	46.3	3.0	0.0	No Failure
3	3.0	L47182	L	298.1	308.5	1498.0	49.4	5.0	0.0	No Failure
4	4.0	L47183	L	298.2	308.6	1433.0	39.5	7.0	0.0	No Failure
5	5.0	L47184	L	298.2	308.7	1408.0	40.0	9.0	0.0	No Failure
6	6.0	M14865	M	298.1	308.6	1425.0	41.9	11.0	0.0	No Failure
7	7.0	L47186	L	298.1	308.6	1558.0	42.4	14.0	0.0	No Failure
8	8.0	L47187	L	298.1	308.6	1527.0	40.2	16.0	0.0	No Failure
9	9.0	M14868	M	298.3	308.7	1667.0	28.6	18.0	0.0	No Failure
10	10.0	M14869	M	298.5	309.0	1741.0	28.0	21.0	0.0	No Failure
11	11.0	H29424	H	298.4	308.9	1782.0	23.9	24.0	0.0	No Failure
12	12.0	H29425	H	298.6	309.1	1423.0	44.3	29.0	0.0	No Failure
13	13.0	M14872	M	298.6	309.1	1339.0	51.1	34.0	0.0	No Failure
14	14.0	M14873	M	298.6	309.2	1742.0	30.0	37.0	0.0	No Failure
15	15.0	L47194	L	298.6	309.2	2035.0	19.6	40.0	0.0	No Failure
16	16.0	L47195	L	298.6	309.2	1542.0	48.4	42.0	0.0	No Failure
17	17.0	M14876	M	298.6	309.2	1311.0	46.6	44.0	0.0	No Failure
18	18.0	M14877	M	298.7	309.2	1410.0	45.6	47.0	0.0	No Failure
19	19.0	H29432	H	298.8	309.2	1306.0	54.5	50.0	0.0	No Failure
20	20.0	M14879	M	298.9	309.3	1632.0	32.5	55.0	0.0	No Failure
21	21.0	H29434	H	298.9	309.3	1375.0	42.7	58.0	0.0	No Failure
22	22.0	L47201	L	298.8	309.3	1450.0	44.8	63.0	0.0	No Failure
23	23.0	M14882	M	298.9	309.3	1581.0	30.7	65.0	0.0	No Failure
24	24.0	L47203	L	299.0	309.4	1758.0	25.7	68.0	0.0	No Failure
25	25.0	M14884	M	299.0	309.4	1561.0	37.3	70.0	0.0	No Failure
26	26.0	L47205	L	299.0	309.5	1861.0	23.3	73.0	0.0	No Failure
27	27.0	L47206	L	299.1	309.5	1512.0	39.0	75.0	0.0	No Failure
28	28.0	H29441	H	299.1	309.4	1811.0	24.6	77.0	0.0	No Failure
29	29.0	L47208	L	299.1	309.4	1439.0	44.2	82.0	0.0	No Failure
30	30.0	L47209	L	299.0	309.4	1693.0	30.1	84.0	0.0	No Failure
31	31.0	M14890	M	299.1	309.5	1339.0	48.2	86.0	0.0	No Failure

As you can see, from my 1st source, there's no need to clean and modify the dataset. There you can see the minimal records, format of each data, units, and the range of values. Down below are my statistics for this dataset.



Filter

Choose None

Current relation

Relation: predictive_maintenance

Instances: 10000

Attributes

All None

No.	
1	UDI
2	Product ID
3	Type
4	Air temperature [K]
5	Process temperature [K]
6	Rotational speed [rpm]
7	Torque [Nm]
8	Tool wear [min]
9	Target
10	Failure Type

Status

OK