ENPM808Y - Fundamentals of AI and Machine Learning

The goal of this assignment is to understand and implement different types of regression and classification algorithms using Python. The specific algorithms to be used are:

Algorithms: Linear Regression Logistic Regression and Naive Bayes

Tasks:

- Perform EDA on the dataset and pre-processing if necessary.
- Train a linear regression , logistic regression and naive bayes model using the dataset "AI4I 2020 Predictive Maintenance"
- Evaluate the model's performance using Accuracy,Precision,Recall and F1 Score.
- Use the trained model to make predictions on unseen data.
- Compare the performance of Linear Regression, Logistic Regression and Naive Bayes.
- Analyze the results and draw conclusions.

Dataset Description

The AI4I 2020 Predictive Maintenance Dataset is a synthetic dataset that reflects real predictive maintenance data encountered in industry. The dataset consists of 10000 data points stored as rows with 14 features in columns. The features are:

- **UID:** unique identifier ranging from 1 to 10000
- **product ID**: consisting of a letter L, M, or H for low (50% of all products), medium (30%) and high (20%) as product quality variants and a variant-specific serial number
- **air temperature [K]**: generated using a random walk process later normalized to a standard deviation of 2 K around 300 K
- **process temperature [K]**: generated using a random walk process normalized to a standard deviation of 1 K, added to the air temperature plus 10 K.
- **rotational speed [rpm]:** calculated from a power of 2860 W, overlaid with a normally distributed noise
- **torque [Nm]:** torque values are normally distributed around 40 Nm with a f = 10 Nm and no negative values.
- **tool wear [min]:** The quality variants H/M/L add 5/3/2 minutes of tool wear to the used tool in the process.
- The output feature is 'machine failure' label that indicates whether the machine has failed in this particular datapoint for any of the following failure modes are true.

  The machine failure consists of five independent failure modes :

- **Tool wear failure (TWF):** the tool will be replaced or fail at a randomly selected tool wear time between 200-240 mins (120 times in our dataset). At this point in time, the tool is replaced 69 times, and fails 51 times (randomly assigned).
- **heat dissipation failure (HDF):** heat dissipation causes a process failure, if the difference between air- and process temperature is below 8.6 K and the tools rotational speed is below 1380 rpm. This is the case for 115 data points.
- **power failure (PWF)**: the product of torque and rotational speed (in rad/s) equals the power required for the process. If this power is below 3500 W or above 9000 W, the process fails, which is the case 95 times in our dataset.
- **overstrain failure (OSF)**: if the product of tool wear and torque exceeds 11,000 minNm for the L product variant (12,000 M, 13,000 H), the process fails due to overstrain. This is true for 98 data points.

- **random failures (RNF)**: each process has a chance of 0,1 % to fail regardless of its process parameters. This is the case for only 5 data points, less than could be expected for 10

Deliverable:

- A report (one page) that includes:
    - Introduction (brief explanation of the problem, dataset, and algorithms used)
    - EDA (data visualization and pre-processing)
    - Model implementation and evaluation
    - Conclusion (summary of the results and any insights you gained)
- All the code used in the analysis.

Note:

- Make sure to use the latest version of Python, and you should use libraries such as pandas, numpy, matplotlib,sklearn and seaborn.
- Use appropriate evaluation metrics for the type of problem you are solving and the algorithm you are using.
- Make sure to explain your thought process and any assumptions you make clearly in the report.