Introduction

Problem: Given on a dataset we have to implement 2 different types of regression and classification algorithms using Python. The specific algorithms to be used are: Logistic Regression, Naïve Bays.

Logistic Regression: Logistic regression is a machine learning algorithm used for classification problems, where the goal is to predict the outcome of a dependent variable based on previous observations. The output of is a probability value between 0 and 1, which represents the likelihood of the input belonging to a certain class. The model learns to assign higher probabilities to inputs that are more likely to belong to the positive class, and lower probabilities to inputs that are more likely to belong to the negative class.

$$1 / (1 + \exp(-z))$$

Naïve Bayes: Naive Bayes is another machine learning algorithm that is used for classification problems. It is based on Bayes' theorem, which states that the probability of a hypothesis (in this case, a class) given some evidence (in this case, input features) is proportional to the likelihood of the evidence given the hypothesis, multiplied by the prior probability of the hypothesis. Naive Bayes assumes that the input features are conditionally independent given the class label, which means that the presence or absence of a particular feature does not depend on the presence or absence of any other feature. The algorithm works by computing the posterior probability of each class given the input features, using Bayes' theorem:

$$P(A|B) = P(A|B) * P(B) / P(A)$$

Dataset: The AI4I 2020 Predictive Maintenance Dataset is a synthetic dataset that reflects real predictive maintenance data encountered in industry. Since real predictive maintenance datasets are generally difficult to obtain and in particular difficult to publish, we present and provide a synthetic dataset that reflects real predictive maintenance encountered in industry to the best of our knowledge.

EDA: In data visualization we see that the dataset consists of 10000 data points stored as rows with 14 features in columns and they are in different datatypes. For accuracy of training model, we check and drop duplicate values from the dataset. Also we remove non-numeric data columns (UID, Product ID) and convert all columns to float type.

Model Implementation: We have implemented 2 different models on this dataset. Logistic Regression model and Gaussian Naïve Bayes Model. Each of them have their individual mathematical equation. Python sklearn was used to perform training and implementing mathematical equations for the models. Error of the model is determined by mean squared error function. For the confusion matrix of the model I had to use ConfusionMatrixDisplay.from_estimator(lr, X_test, y_test) instead of plot_confusion_matrix(lr, X_test, y_test) as it seems that plot_confusion_matrix is deprecated. My IDE was not accepting import for plot_confusion_matrix.

Conclusion: After implementation of this model and from the test data our training accuracy stands as following:

	Accuracy	Training Accuracy
Logistic Regression	61.270000	59.460000
Naïve Bayes	30.430000	31.900000