ESI 4606: Analytics I - Foundations of Data Science
Mid-term Exam - Part I
Due: October 26st (11:30PM), 2022

R Programming Problem (40 points)

Considering a sample of electronic product units from a manufacturing company running under different levels of operating conditions (i.e., temperature and voltage) with different material suppliers (i.e., suppliers A and B), the corresponding success-failure data within the mission time period has been collected. Two data sets are included, namely a **training data set** of 200 product units (in "data.train2022.txt") and a **test data set** of 300 product units (in "data.test2022.txt"). Variable description of the data sets is given in Table 1. For instance, a data record of "failure=N, temp=2.88, stress=1, material=1" means that a product unit using material provided by supplier B, operating at temperature of 2.88 Celsius degree and high voltage survives within the mission time period.

Table 1: Variable description

Variable	**Variable type**	**Description**
failure	categorical variable	Y: failure; N: success
temp	numerical variable	temperature values in degrees of Celsius
stress	dummy variable	0: low voltage (coded as baseline); 1: high voltage
material	dummy variable	0: supplier A (coded as baseline); 1: supplier B

Using variable "failure" as the response variable and the rest of variables as input variables, **please answer all questions below clearly and include R codes in the appendix sections**. After importing both data sets, please change response variable "failure" in each data set as a factor object.

(1) **(4 points)** Based on the training data, draw side-by-side boxplots of product operating temperature values under two product groups, namely failure group and success group. What findings can you draw from such data visualization regarding the influence of the operating temperature condition on the occurrence of product failure?

(2) **(4 points)** Fit Linear Discriminant Analysis (LDA) model using the training data. Compute the confusion matrices and the mis-classification error rates for both the training data and test data, respectively.

(3) **(4 points)** Repeat (2) and answer the same questions using Quadratic Discriminant Analysis (QDA).

(4) **(4 points)** Repeat (2) and answer the same questions using Logistic Regression.

(5) **(4 points)** Repeat (2) and answer the same questions using K-Nearest Neighbor (KNN) classification with $K = 2$. (Note: set your random seed for KNN as 2022)

(6) **(4 points)** Repeat (2) and answer the same questions using KNN classification with $K = 7$. (Note: set your random seed for KNN as 2022)

(7) **(5 points)** Repeat (2) and answer the same questions using Multiple Linear Regression. (Hint: Since response is categorical variable, you need to first numerically convert "Y" and "N" into 1 and 0 respectively, before running multiple linear regression. After prediction, you need to convert the predicted values into "Y" and "N" before computing the confusion matrix and the mis-classification error rate, according to the following rule: if the predicted value is larger than or equal to 0.5, convert it into "Y"; if the predicted value is smaller than 0.5, convert it into "N".)

(8) **(6 points)** If the main objective of this study is to develop a data-driven model for better prediction of the occurrence of product failure so that the sub-subsequent proactive maintenance decisions can be better informed, based on results from (2)-(7), which model, do you think, will be the most appropriate and explain why? Which model, do you think, will be the least appropriate and why it does not perform well as compared to other models?

(9) **(5 points)** If the main objective of this study is to develop a data-driven model for better understanding and quantification of different factors' (i.e., temperature, material supplier, voltage) influence on the occurrence of product failure so that R&D decisions for reliability improvement can be better informed, based on results from (2)-(7), which model, do you think, will be the most appropriate and explain why? Based on this model, further interpret how different factors influence the occurrence of product failure.