**Data Science HW3 (Due on 11/05 24:00)**

1. Based on pima dataset, remove the samples with zeros or missing values and treat “Outcome” as the label. Construct logit model identify the significant predictors. Based on these significant predictors, conduct classification using KNN (*K* value needs to be optimized) and Naïve Bayes. Please compare their performances in terms of accuracies, recall, precision, and *F* score.
2. Based on the winequality-red dataset, applying MLR to select the significant features (response is quality and *P* value is less than 5%). Then, based on these features, conduct regression using KNN (*K* value needs to be optimized) and MARS (degree & penalty need to be optimized). Compare their performances in terms of RMSE, MAE, and MAPE.
3. Based on bank dataset, remove Job, Month, Day and treat “y” as the label: (a) construct logit model to identify the significant predictors (*P* value is less than 5%). (b) derive the probability respectively for y-yes or y-no by including a new sample: discrete features are just the mode of all samples and numeric features are just the median of all samples. (c) apply Naïve Bayes to conduct classification with regard to using the partial predictors identified by logit model and using all predictors (in terms of overall accuracy and respective accuracy for y-yes and y-no).
4. Based on container dataset (response is Container), applying MLR and MARS to identify the key performance indicators (using the union of the identified indicators as input features). Managerial insights are required to provide. Compare the performances in term of RMSE, MAE, and MAPE. Draw a plot to demonstrate the actual value and the predictive results. Remember to split the dataset into training (2011~2020) and testing (2021~2023) before constructing models.
5. Redo the Titanic dataset and perform the following comparisons between Naïve Bayes and Logit model (70% for training and 30% for testing): (1) compare the impact of gender (male vs female) given age=30, class 2, fare=20, (2) compare the impact of age (1 vs 25) given gender=male, class 1, fare=40, and (3) compare the impact of class (1, 2 ,3) given gender=female, age=25, fare=40. Indicate the differences between these two methods.
6. Based on the handysize dataset (response is Handysize), apply Ridge, Lasso, and ElasticNet to predict aggregate sales. What significant predictors are identified? Compare their performances in terms of RMSE, MAE, and MAPE. Remember to remove data samples with missing values. Draw a plot to demonstrate the actual value and the predictive results (no need to split the dataset into training and testing).