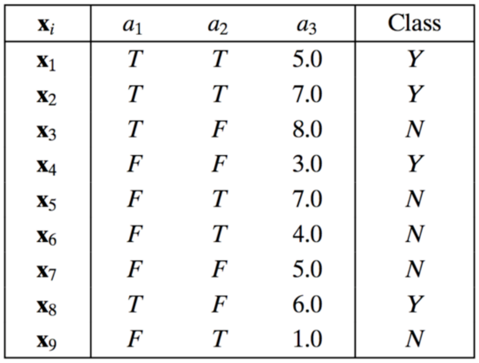
IE 7275 Data Mining in Engineering
Homework 2
Deadline: 10/14/2021

Note:

- **Submit all necessary files along with your solution sheets on Canvas.**
- **You use Python or equivalent programming language only.**

Problem 1 (Classification via Bayes' Classifier) [10 points]

Given the dataset in Table 1, build a naïve Bayes classifier to classify the new point (T, F, 7). Show all steps in your calculation. Hint: independency between variables.



$\mathbf{x}_i$	a_1	a_2	a_3	Class
$\mathbf{x}_1$	*T*	*T*	5.0	*Y*
$\mathbf{x}_2$	*T*	*T*	7.0	*Y*
$\mathbf{x}_3$	*T*	*F*	8.0	*N*
$\mathbf{x}_4$	*F*	*F*	3.0	*Y*
$\mathbf{x}_5$	*F*	*T*	7.0	*N*
$\mathbf{x}_6$	*F*	*T*	4.0	*N*
$\mathbf{x}_7$	*F*	*F*	5.0	*N*
$\mathbf{x}_8$	*T*	*F*	6.0	*Y*
$\mathbf{x}_9$	*F*	*T*	1.0	*N*

Problem 2 (Bayes and KNN Classification) [25 points]

(a) Consider the admission dataset (training and testing subtest), build a naïve Bayes classifier with the training dataset (assuming data are normally distributed), and then make a prediction for testing dataset. Construct two confusion matrices for training and testing datasets and compute *accuracy*.

(b) Build a KNN classifier with the training dataset, and then make a prediction for testing dataset. Construct two confusion matrices for training and testing datasets and compute *accuracy*. Specify *k* and distance measurement used in your classifier.

(c) Compare the results between parts (a) and (b).

Problem 3 (KNN Classification) [15 points]

See Problem 7.3 in the textbook: Predicting Housing Median Price

Problem 4 (Bayes Classification) [15 points]

See Problem 8.1 in the textbook: Personal Loan Acceptance

The dataset contains 1500 applications.

Problem 5 (Performance Evaluation) [10 points]
See Problem 5.6 in the textbook.

Problem 6 (Performance Evaluation) [10 points]
See Problem 5.7 in the textbook.

Problem 7 (Performance Evaluation for Bayes Classification) [15 points]
Implement a Bayes Classification for the dataset RidingMowers. The probability for target class "owner" is computed as follows: P(C_owner|X)/P(C_owner|X) + P(C_non-owner|X). Construct a ROC curve by considering cut-off values {0, 0.25, 0.5, 0.75, 1} and compute the AUC.