Assignment 4

Note: Show all your work.

Problem 1 (20 points) Consider the following confusion matrix.

	predicted class		
actual class		C1 (positive)	C2 (negative)
	C1 (positive)	254	36
	C2 (negative)	72	324

Compute *sensitivity*, *specificity*, *precision*, *accuracy*, *F-meassure*, *F*2, and MCC measures. You have to show all your calculations.

Problem 2 (20 points) Suppose you built two classifier models *M*1 and *M*2 from the same training dataset and tested them on the same test dataset using 10-fold cross-validation. The error rates obtained over 10 iterations (in each iteration the same training and test partitions were used for both *M*1 and *M*2) are given in the table below. Determine whether there is a significant difference between the two models using the statistical method discussed that we discussed in the class (this method is also discussed in Section 8.5.5, pp 372-373 of the textbook). Use a significance level of 1%. If there is a significant difference, which one is better?

Iteration	M1	M2
1	0.12	0.08
2	0.12	0.1
3	0.15	0.22
4	0.15	0.1
5	0.03	0.07
6	0.17	0.11
7	0.2	0.1
8	0.14	0.11
9	0.1	0.17
10	0.14	0.11

Note: When you calculate *var(M1 – M2)*, calculate a sample variance (not a population variance).

You must show all calculations, including the calculation of the test statistic.

Problem 3 (20 points). The following table shows a test result of a classifier on a dataset.

Tuple_id	Actual Class	Probability
1	P	0.92
2	N	0.70
3	N	0.76
4	P	0.92
5	P	0.83

6	P	0.89
7	N	0.79
8	P	0.73
9	N	0.82
10	P	0.96

(1). For each row, compute *TP*, *FP*, *TN*, *FN*, *TPR*, and *FPR*.
(2). Plot the ROC curve for the dataset. You must draw the curve yourself (i.e., don't use Weka, R, or other software to generate the curve).

Problem 4 (20 points). This problem is a small experiment of handling an unbalanced dataset for classification. Use *a3_p4_train.arff* and *a3_p4_test.arff* files and use J48 on Weka.

(1). Build a decision tree model from *a3_p4_train.arff* using J48 and test it on *a3_p4_test.arff*. Include the resulting confusion matrix in your submission.
(2). Create an undersampled training dataset from *a3_p4_train.arff* and name it *a3_p4_train_undersampled.arff*. Build a decision tree model from *a3_p4_train_undersampled.arff* using J48 and test it on *a3_p4_test.arff*. Include the resulting confusion matrix in your submission.
(3). Create an oversampled training dataset from *a3_p4_train.arff* and name it *a3_p4_train_oversampled.arff*. Build a decision tree model from *a3_p4_train_oversampled.arff* using J48 and test it on *a3_p4_test.arff*. Include the resulting confusion matrix in your submission.
(4). What conclusion can you draw from this experiment?

Note:

- You may use any tool(s) when creating undersampled dataset and oversampled dataset. You must describe the tool(s) you used.
- If you know how to use Python for classification, you may use Python's *DecisionTreeClassifier* instead of Weka's J48. In this case, you must submit Python script file(s) as well as confusion matrices. Use *a3_p4_train.csv* and *a3_p4_test.csv* files

Problem 5 (20 points). Use JMP Pro to build and test five classifier models – Naïve Bayes, KNN, Partition (decision tree), Boosted Tree, and Neural Network – following the instruction in *JMP-classification-assignment.pdf* file.

Submission:

Include all answers in a single file and name it *LastName_FirstName*_HW4.EXT. Here, "EXT" is an appropriate file extension (e.g., docx or pdf). If you have multiple files, then combine all files into a single archive file. Name the archive file as *LastName_FirstName*_HW4.EXT. Here, "EXT" is an appropriate archive file extension (e.g., zip or rar). Upload the file to Blackboard.