AMERICAN INTERNATIONAL UNIVERSITY-BANGLADESH

Choose an item.

Assignment Cover Sheet

Assignment Title:	Project – Supervised Learning			
Assignment No:	01		Date of Submission:	15 May 2020
Course Title:	Data Warehousing and Data Mining			
Course Code:	CSC4139		Section:	A
Semester:	Spring	2019-20	Course Teacher:	Rahman Mohammod Hafizur

Declaration and Statement of Authorship:

1. I/we hold a copy of this Assignment/Case-Study, which can be produced if the original is lost/damaged.
2. This Assignment/Case-Study is my/our original work and no part of it has been copied from any other student's work or from any other source except where due acknowledgement is made.
3. No part of this Assignment/Case-Study has been written for me/us by any other person except where such collaborationhas been authorized by the concerned teacher and is clearly acknowledged in the assignment.
4. I/we have not previously submitted or currently submitting this work for any other course/unit.
5. This work may be reproduced, communicated, compared and archived for the purpose of detecting plagiarism.
6. I/we give permission for a copy of my/our marked work to be retained by the Faculty for review and comparison, including review by external examiners.
7. I/we understand thatPlagiarism is the presentation of the work, idea or creation of another person as though it is your own. It is a formofcheatingandisaveryseriousacademicoffencethatmayleadtoexpulsionfromtheUniversity. Plagiarized material can be drawn from, and presented in, written, graphic and visual form, including electronic data, and oral presentations. Plagiarism occurs when the origin of them arterial used is not appropriately cited.
8. I/we also understand that enabling plagiarism is the act of assisting or allowing another person to plagiarize or to copy my/our work.

* *Student(s) must complete all details except the faculty use part.*
** Please submit all assignments to your course teacher or the office of the concerned teacher.

Group Name/No.:

No	Name	ID	Program	Signature
1	Ahmed, Sinthia	17-33820-1	BSc [CSSE]	
2			Choose an item.	

Faculty use only		
FACULTYCOMMENTS	**Marks Obtained**	
	Total Marks	

Project- Supervised Learning

Problem Title - Teaching Assistant Evaluation

Problem Definition –

The data consist of evaluations of teaching performance over three regular semesters and two summer semesters of 151 teaching assistant (TA) assignments at the Statistics Department of the University of Wisconsin-Madison. The scores were divided into 3 roughly equal-sized categories ("low", "medium", and "high") to form the class variable.

Objective: Objective is to find out best classifier from 5 different classifiers to classify teaching performance based on the Teaching Assistant Evaluation Dataset using Weka tool.

Number of Instances: 151

Number of Attributes: 6 (including the class attribute)

Attribute Information:

1. Whether of not the TA is a native English speaker (binary)
   1=English speaker, 2=non-English speaker
2. Course instructor (categorical, 25 categories)
3. Course (categorical, 26 categories)
4. Summer or regular semester (binary) 1=Summer, 2=Regular
5. Class size (numerical)
6. Class attribute (categorical) 1=Low, 2=Medium, 3=High

Missing Attribute Values: None

Preparation of Dataset –

The dataset was downloaded from UCI repository (http://archive.ics.uci.edu/ml/datasets/Teaching+Assistant+Evaluation). There were no missing attribute values so the dataset was converted into .arff file and loaded into weka tool for further use.

Tested Classifiers –

1. Naïve Bayes
2. IBK
3. KStar
4. J48
5. Random Tree

Naive Bayes Classifier:

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances	76	50.3311 %
Incorrectly Classified Instances	75	49.6689 %
Kappa statistic	0.2547	
Mean absolute error	0.3732	
Root mean squared error	0.4637	
Relative absolute error	83.9862 %	
Root relative squared error	98.3569 %	
Total Number of Instances	151	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.469	0.265	0.460	0.469	0.465	0.204	0.680	0.548	1
	0.460	0.248	0.479	0.460	0.469	0.215	0.663	0.459	2
	0.577	0.232	0.566	0.577	0.571	0.343	0.699	0.559	3
Weighted Avg.	0.503	0.248	0.503	0.503	0.503	0.255	0.681	0.523	

=== Confusion Matrix ===

a	b	c	<-- classified as
23	14	12	a = 1
16	23	11	b = 2
11	11	30	c = 3

IBK Classifier:

IB1 instance-based classifier
using 3 nearest neighbour(s) for classification for better Results

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances	59	39.0728 %
Incorrectly Classified Instances	92	60.9272 %
Kappa statistic	0.0873	
Mean absolute error	0.3703	
Root mean squared error	0.4988	
Relative absolute error	83.3304 %	
Root relative squared error	105.8064 %	
Total Number of Instances	151	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class

	0.449	0.431	0.333	0.449	0.383	0.017	0.600	0.451	1
	0.240	0.248	0.324	0.240	0.276	-0.008	0.605	0.459	2
	0.481	0.232	0.521	0.481	0.500	0.254	0.708	0.497	3
Weighted Avg.	0.391	0.302	0.395	0.391	0.388	0.090	0.639	0.470	

=== Confusion Matrix ===

a	b	c	<-- classified as
22	16	11	a = 1
26	12	12	b = 2
18	9	25	c = 3

KStar Classifier:

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances	94	62.2517 %
Incorrectly Classified Instances	57	37.7483 %
Kappa statistic	0.4349	
Mean absolute error	0.2844	
Root mean squared error	0.4151	
Relative absolute error	63.9945 %	
Root relative squared error	88.0491 %	
Total Number of Instances	151	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.673	0.284	0.532	0.673	0.595	0.370	0.816	0.687	1
	0.580	0.188	0.604	0.580	0.592	0.396	0.766	0.663	2
	0.615	0.091	0.780	0.615	0.688	0.560	0.818	0.680	3
Weighted Avg.	0.623	0.186	0.642	0.623	0.626	0.444	0.800	0.677	

=== Confusion Matrix ===

a	b	c	<-- classified as
33	12	4	a = 1
16	29	5	b = 2
13	7	32	c = 3

J48 Classifier:

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances	80	52.9801 %
Incorrectly Classified Instances	71	47.0199 %
Kappa statistic	0.2948	

Mean absolute error 0.3549
Root mean squared error 0.4793
Relative absolute error 79.8549 %
Root relative squared error 101.6727 %
Total Number of Instances 151

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.551	0.294	0.474	0.551	0.509	0.248	0.681	0.418	1
	0.420	0.198	0.512	0.420	0.462	0.235	0.617	0.463	2
	0.615	0.212	0.604	0.615	0.610	0.401	0.725	0.606	3
Weighted Avg.	0.530	0.234	0.531	0.530	0.528	0.297	0.675	0.498	

=== Confusion Matrix ===

a	b	c	<-- classified as
27	11	11	a = 1
19	21	10	b = 2
11	9	32	c = 3

Random Tree Classifier:

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances 100 66.2252 %
Incorrectly Classified Instances 51 33.7748 %
Kappa statistic 0.4935
Mean absolute error 0.2196
Root mean squared error 0.438
Relative absolute error 49.4072 %
Root relative squared error 92.9102 %
Total Number of Instances 151

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.673	0.186	0.635	0.673	0.653	0.480	0.802	0.638	1
	0.620	0.188	0.620	0.620	0.620	0.432	0.753	0.568	2
	0.692	0.131	0.735	0.692	0.713	0.569	0.811	0.664	3
Weighted Avg.	0.662	0.168	0.664	0.662	0.663	0.495	0.789	0.624	

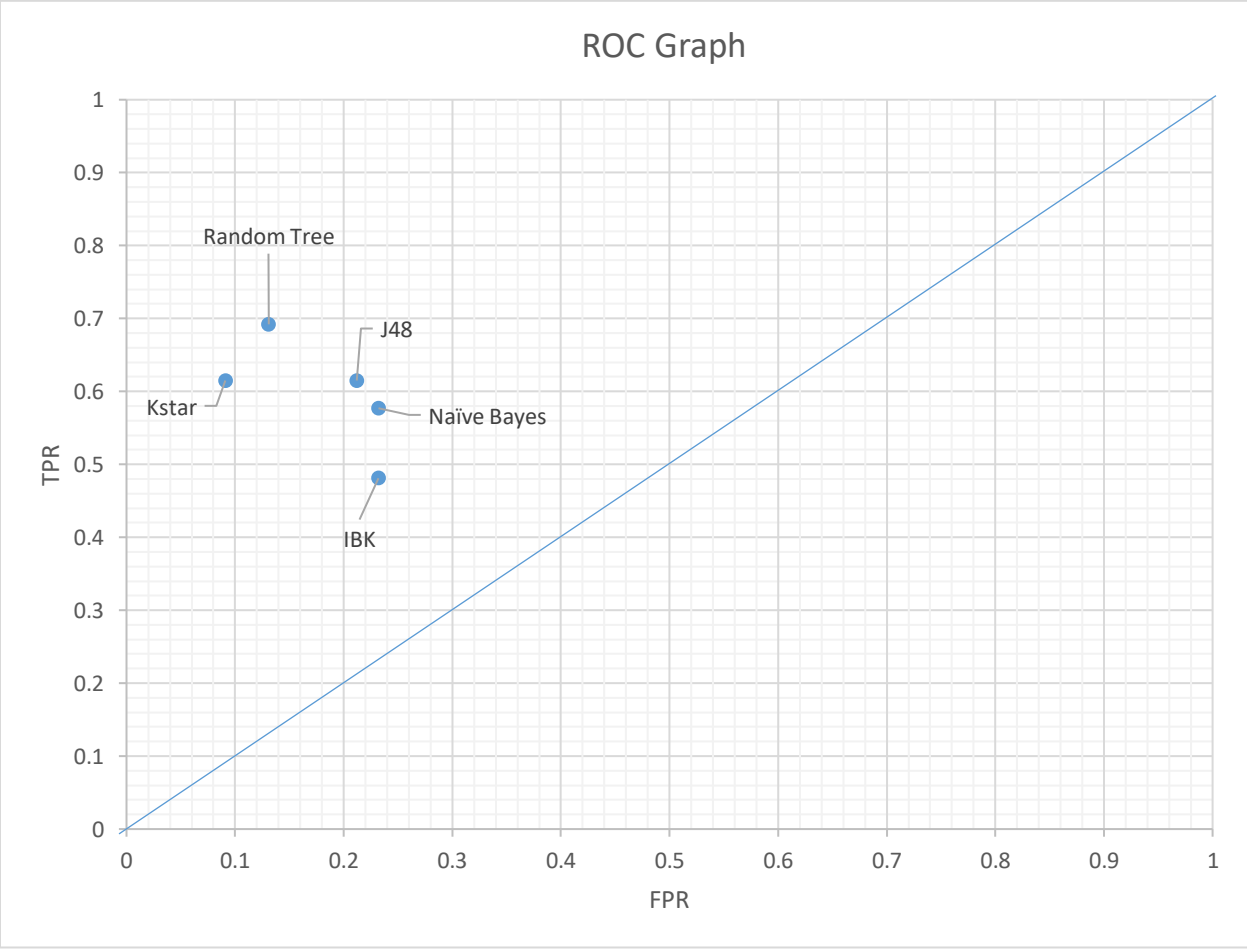
=== Confusion Matrix ===

a	b	c	<-- classified as
33	11	5	a = 1
11	31	8	b = 2
8	8	36	c = 3

Let, C=3(High Performance) be our Positive Interest

Classifier	TPR	FPR
Naïve Bayes	0.577	0.232
IBK	0.481	0.232
Kstar	0.615	0.091
Random Tree	0.692	0.131
J48	0.615	0.212

ROC Graph:



Comment:

The dataset is about Teaching Assistant performance where 1= Low, 2= Medium and 3= High performance. So a classifier that can give more True Positive Value, which means an approximate amount of TA with High performance and low False Positive Value, which means small number of mistakenly classified TA as High performance, will be a good Classifier. From the ROC graph, we can see that all the 5 classifiers are in good region but the Random tree classifier is more close to the best point (0, 1) than Kstar, J48, Naïve Bayes and IBK classifiers. It is also seen that the Random tree classifier's True Positive Rate (0.692) is higher than other classifiers, which means it has more ability to correctly predict High-performance TA than other classifiers. The Random tree classifier's False Positive Rate (0.131) is smaller than J48, Naïve Bayes and IBK classifier's FPR, which means its ablity to mistakenly classify TA with High performance is lower than those classifer. Even though the Random Tree classifier's FPR is larger than Kstar classifier (0.091) we will not consider Kstar, as its TPR is smaller than Random tree. From Random Tree, Kstar, J48, Naïve Bayes and IBK classifiers IBK is the worst classifier because its True Positive Rate is lower and False Positive Rate is higher than other classifiers which makes it unreliable. From studying the ROC graph we can conclude that the Random Tree classifier is the best classifier in this scenario.