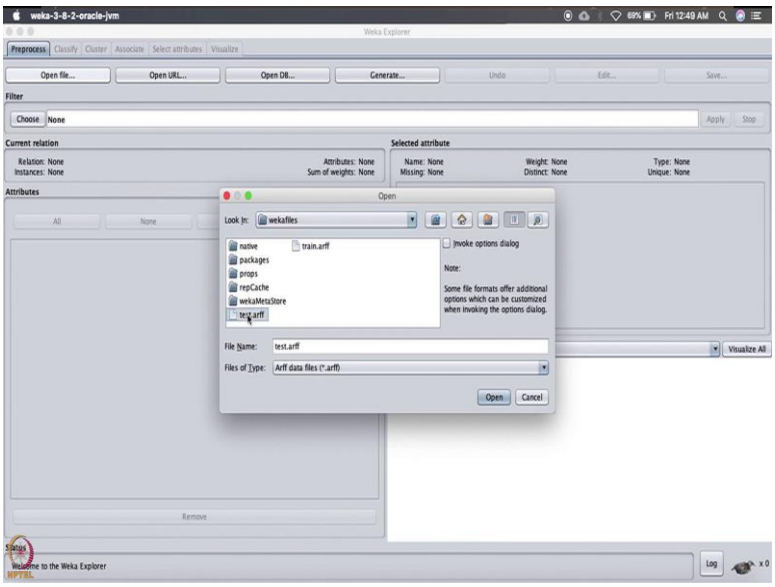
**Introduction to Learning Analytics**
**Prof. Ramkumar Rajendran**
**Department of Interdisciplinary Programme in Educational Technology**
**Indian Institute of Technology, Bombay**

**Lecture - 16**
**Weka demo and how to read the results**

(Refer Slide Time: 00:17)

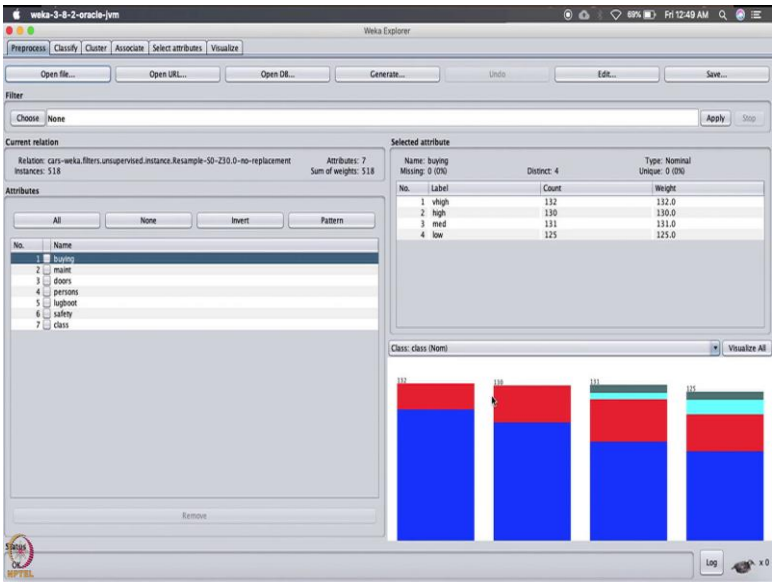



Welcome back to Introduction to Learning Analytics course. In this learning dialogue, we will demonstrate how to use Weka. I hope in last LxT; you might have install Weka and you have already played with Weka. However, in this LED; I will talk about how to use Weka with a simple data set.

We can start Weka in explorer mode, you get this particular screen. The first the top; tabs are the preprocessing that is you can open some data and you can store the data and the tabs clusters, classification and the associate like identifying the rules association rules between the data; then also the tab call select attributes. This is to select if you have a lot of independent variables; we can select attributes by automatically and also we can visualize the data for the analysis.

You can load the data on Weka using this four different tabs; like Open URL or open file let us do the Open file. We already loaded a data call test.arff; arff is the format which is used for the Weka.

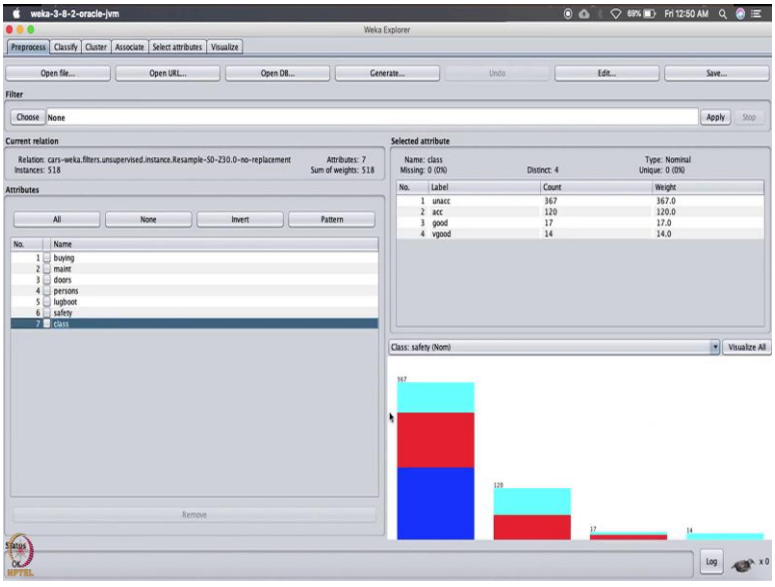
(Refer Slide Time: 01:39)




Now, this data is actually cars data that is there are lot of features in the car; there are like 7 features; 6 features and 7th is the label you want to predict. And the 518 instances of this features; we have data from 518 different cars and there are 7 features.

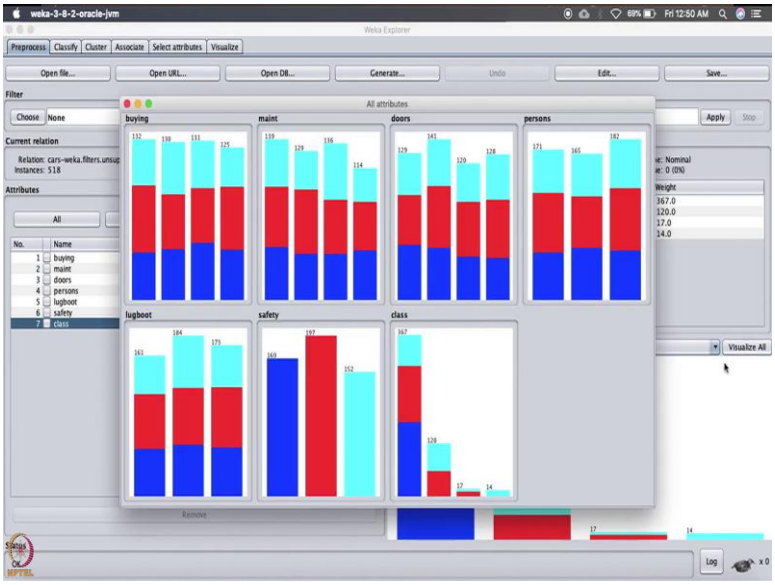
So, let us see the feature 1 that is buying and there is no missing data; we have all the data that is good and buying as a four different values this like four different values; it is a nominal type data the values are high, very high, high, medium and low. Now, this data is compared with a class called class like that is a label you have predicted.

(Refer Slide Time: 02:41)



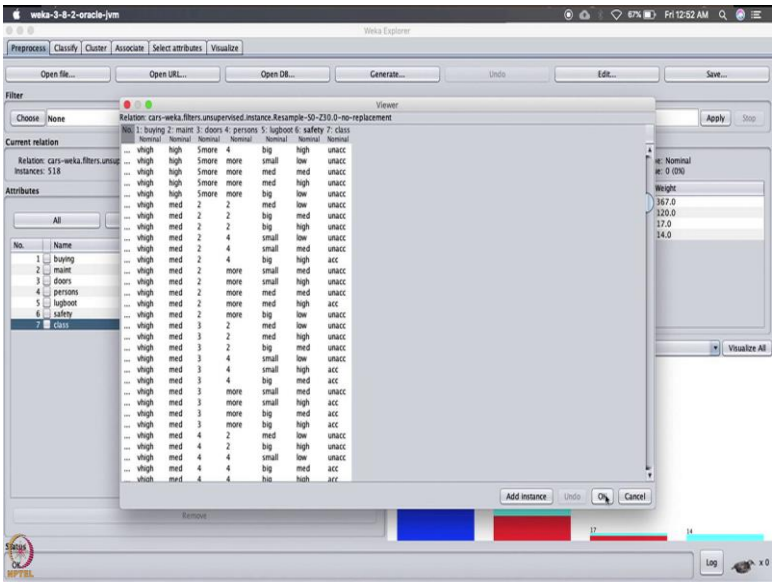
So, if we can compare; visualize this data across different other variables; for example I am plotting the buying versus safety here. Just look at the maintenance; maintenance also have similar variables like very high something. And number of doors in the car is 2, 3, 4, 5 more; number of persons can sit in the cars like 2, 4 or more like a space safety is high low medium and the class.

(Refer Slide Time: 03:05)




If you look at the class or the safety across all this variables; here we are plotting the variable buying a maintenance doors, persons, luggage space versus safety. The safety is three variables that is medium, high, low; there are like a 169 instances of low, 197 instances of medium and 152 instances of high.

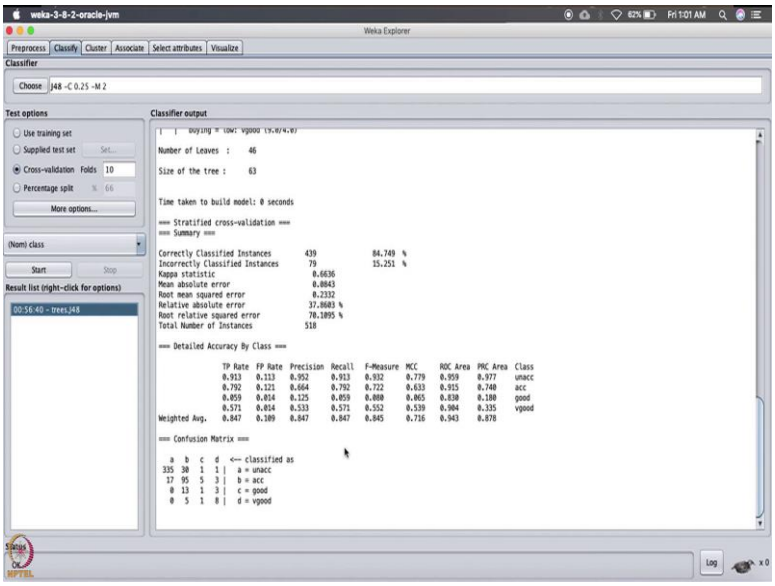
(Refer Slide Time: 03:35)



You can edit the data like if it has any missing values; in this particular data we do not have any missing value, if we has any missing values we can edit the data. So, all these data is nominal. So, we can see the values for each of these variables like a buying, maintenance, doors, persons, luggage space, safety and classes. There are four classes in this; there are four classes in this variable, one is unacceptable, acceptable condition, good condition and very good condition. The aim of this data set is you have a 518 cars data and you have to predict whether the car; given the data the car can be classified as unacceptable or acceptable or very good condition or good condition; there are seven data sets here.

So, first one is buying like the chance of buying is high very high, maintenance is good very high medium not maintainable. And the numbers of doors and persons and luggage space and safety; the safety is 3, 3 features that is low, medium and high. Given this six features we have to predict whether the car is classified as acceptable or unacceptable or good or very good. So, these are the data we obtained from 518 cars; also these class of classification is also done by the manually the experts in the automobile field.

(Refer Slide Time: 05:31)




The next important tab which we will do is a classify; you will not deal you will not talk about other tabs in this demonstration. In a Weka we can the default classifier is zero that is a base line classifier; here first we will start with the J48 classifier in a decision trees; decision tree is the another algorithm to classify the given variable into four classes.

Now, you have four different text options; one is you can use the training set that is the 518 data we used to train the model, use the same data set to test it. If you do that; you get always highest accuracy or you can supply the separate test set; like uses this 518 instance of data to create the model; then supply another set of data to test it. Or you can split the data by percentage wise; say I want use 66 percentage; that is two third of data to train this model and use the other one third of data to test the model.

However, there is a bias which two third of model you are going to split it or you randomly picking this two third of data or this two third of data is a first 66 instance of data. In order to avoid this confusions, we can use the cross validation in a cross validation we split the given data set into set of wholes; multiple groups. Say if I want to I have a data; I want to do the 10 fold cross validation you have to split the data into 10 different groups.

Let us take it very simple cross fold validation example you have a 30 instance of data and you want to do three cross three fold cross validation. You can split the data into first 10 is a first fold and second 10 data in second group and third 10 data in third groups. So,

that three groups of data each have 10; 10 instance of data. What we will do? We will train the first 20 data that is group 1 and group 2 data to create the model we test it on the group 3 data; so the group 3 data is a test data. Now, in a next iteration we select the group 1 and group 3 as the training data set with create a model using group 1 and group 3 as a training set; then we will test it using the group 2; so the group 2 also tested now.

In a next iteration, we will select group 2 and group 3 as a training model then we will select group 1 for the testing. So, in this case all this group 1, group 2, group 3 as being used to test the model and the performance of all this group is indicated as the accuracy or the performance of this particular classifier model. So, 10 fold cross validation or the n fold cross validation is very very useful in order to do the classification; please use the cross validation technique when you apply the Weka in your course project.

You can go and look at the more options like instead of just selecting this four given options; we can have more options to high penalizing particular variable or you can you what are the things you want to look at that you want to output the model or you want to output only the performance; all these things can be done.

So, we selected J48 as the classifier model, then we want to predict the class the normally class of acceptable unacceptable very good and good and you start it. Once we start it, we can see the performance here; the first thing is correctly classified instances that is out of 518 instances; 439 instances are correctly classified. So, 439 divided by 518 that is the 83 percentage; 84 percentage that is the accuracy of this particular classifier.

So, incorrectly classified instances is 79 instances are incorrectly classified; then Kappa statistic is 0.6 is very good value. Kappa actually compare the classify performance compared to the 0 or classify that you have the baseline classifier that is if I classify all this values as a; a only that is unacceptable condition what happens that is the condition here in Kappa.

So, this particular is 66, 0.66 better than the baseline classifier. So, 0.6 is a very good value actually in classification, but in education settings it is good, but in this the setting which we views that is example of car classification is not; it may not be better Kappa value. Root mean square is a 0.2332 and this absolute error we can identify; let us look at the table detailed accuracy by class.

So, there is a true positive rate, false positive rate, precision, recall; let us below look at the confusion matrix there are four values a, b, c, d; a is unacceptable, b is acceptable, c is good and d is very good. This a, b, c, d classified as is the predicted value; the actual values given in the row wise the column wise is the predicted value.

There are 335 cars classified as unacceptable which is actually unacceptable that is good. There are 17 cars classified as unacceptable, but they are actually acceptable that is b equal to acceptable. So, there is a wrong classification of 17 data; that is like unacceptable sorry the acceptable cars or classified as a unacceptable value.

So, this is used to calculate the precision of this classifier; if you look at the precision of the first row is a 0.952. So, this value is good also this; the there are 32 other cars which are unacceptable, but which are not able to not classified as a unacceptable by the system. So, you can look at it the 335 cars classified as unacceptable is correct then 30 and 1 and 1; 32 cars which are actually unacceptable, but are classified as acceptable and good and very good.

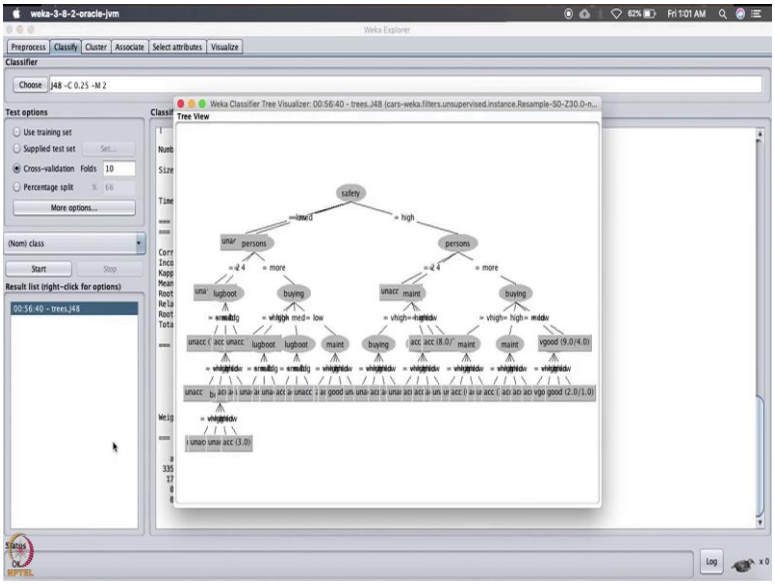
So, the recall rate is reduced to 0.913; so the F measure actually computes based on precision and recall then the values given here. So, you can look at the confusion matrix in a Wikipedia and look at the precision recall what is F measure and ROC area; it will be there will be explain it using this similar kind of table. The interesting thing here is the b acceptable to acceptable 95; it is good. Let us look at the poor performance that is c; there are 13 plus 1 plus 1; that is 17 cars which are good really good. But our model predicted only 1 car is a good c equal to 1 and other cars are acceptable and other 13 cars are acceptable, 3 cars as a very good.

So, this is very poor; this is very poor recall rate also very poor precision rate. So, you will look at the table the third line; you can see the very poor recall rate and precision rate at 0.059 and 0.125; this is very very poor. And the fourth one although it is very less numbers, but it is good; in the sense out of 13 cars sorry 14 cars, 8 plus 1 plus 5; 8 cars are classified as very good. So, this good; so we might have a better recall rate compare to the other ones.

So, in based on this particular classification we can see that the classification of good car is not done (Refer Time: 14:10) good. So, our work compared to all other three classes this classifier has done a really good job; so the Kappa values 0.66. If you train the

different set of classifier, we might able to increase the performance of this classification task.

(Refer Slide Time: 14:33)




You can also view the; the model we created because as decision trees we can visualize the tree. For example the visualize decision trees has been created based on safety; safety is the primary parameter here and if the safety is equal to high, then it checks whether the number of persons in the given car then based on that if it is more whether what is the value of buying based on that it checks a maintenance then it goes and classifies based on the value of maintenance acceptance unacceptable.

Or it can which is lower than it again its looks at persons then the luggage space or buying values. So, this is the real model which is used to create this classify the given data set. Thank you for listening to the Weka demo and I hope you will use Weka in a course project. We actually have the data and we explained the data in an next LED use the data and apply it on a Weka and try to predict the students dropout rate in a MOOC.

Thank you.