FoDA CS910/ CS430 COURSEWORK 1

Student ID: (2045986) MSc Computer Science

November 2020

'Abalone Data' - A Brief

The dataset, termed as 'Abalone Data', made up of 4177 instances comprising of 8 features, describes the physical characteristics of abalones, a common name for small to large marine snails. Among the 8 attributes, one of them is Nominal in nature, while the others being numeric. The number of rings feature provided in the dataset is the value to predict: either as a continuous value or as a classification problem. The age of the abalone is predicted through its ring count as ringCount + 1.5 years. The attributes are as follows: **Sex, Length, Diameter, Height, Whole weight, Shucked Weight, Viscera Weight, Shell Weight, Rings**.

The study of the Abalone Dataset and the results of the coursework provided below were performed on Weka. For all models, **a ten-fold cross validation technique was adopted as the Test Mode**.

1 Diameter as a function of Length

A Simple Linear Regression was performed to predict Diameter as a function of Length. The linear regression line takes the form:

$$y = m * x + c \tag{1}$$

where y is the dependent variable and x is the independent variable. The parameters of the model refer to slope and y-intercept of the line, depicted by m and c respectively. The equation of the line obtained relating length and diameter is as follows:

$$Diameter = 0.8155 * Length + -0.0194 \tag{2}$$

The parameters in the above equation, take the values $m = 0.8155$ and $c = -0.0194$. A value of 0.8155 as slope tells us that, for every 1 millimeter(mm) increment in length of an abalone, we can expect a 0.8155mm increase in its diameter, and that the two attributes are positively correlated to each other. Diameter of the abalone in this context refers to the perpendicular width of its shell. The slope coefficient is in accordance with an intuitive understanding of the abalone's physical appearance, i.e., the ratio of length to diameter is 1:0.8155 implying that the abalone shell's oval shape is not violated by this equation.

Furthermore, the model indicates that when the length of an abalone is 0mm, its diameter takes a negative value of -0.0194 mm (negative y-intercept). This interpretation of the y-intercept does not make sense in the real world, as neither a zero length nor negative diameter have any significance in our use case.

The value of the Correlation Coefficient = 0.9868; indicates a strong positive relationship between the attributes.

Here are other performance metrics for Eq. (2) :-

Metrics	Value
Time taken to build the model	0.01s
Mean absolute error	0.0115
Root mean squared error	0.0161
Relative absolute error	14.3808
Root relative squared error	16.1891

2 Predicting Whole weight of abalone

A Multi-Linear Regression Model uses multiple attributes for predicting the dependent variable. Here, we use the different weight attributes in the dataset such as shucked weight, viscera weight and shell weight to model the whole weight of an abalone. The model obtained is:

$$\begin{aligned} WholeWeight = 0.9366 * ShuckedWeight \\ +1.1116 * VisceraWeight \\ +1.253 * ShellWeight - 0.0078 \end{aligned} \tag{3}$$

Weight Attribute	Slope
$x1$ = Shucked Weight	0.9366
$x2$ = Viscera Weight	1.1116
$x3$ = Shell Weight	1.253

The parts of an abalone contribute to its whole weight differently. The whole weight increases the most, by a factor of 1.253, with a unit increase of Shell Weight, keeping other attributes constant, in comparison to the vice versa. The notion that whole weight as sum of part weights is therefore misguided, though intuitive. Let's consider a few instances of the data through the Viewer facility in Weka Tool shown in Fig. (1), to compute the whole weight as sum of part weights and then clarify with the value of Whole weight attribute.

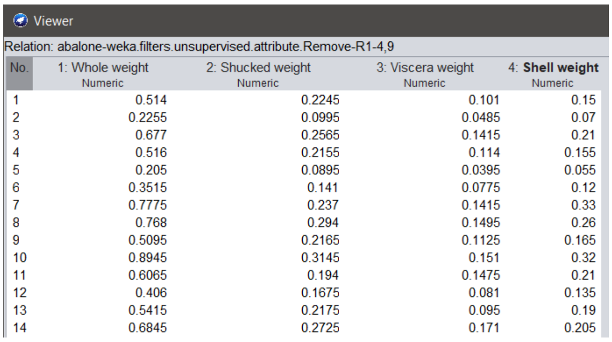
LHS: Shucked Weight + Viscera Weight + Shell Weight

RHS: Whole Weight

Instance 2: **LHS** = 0.2565 + 0.1415 + 0.21 = 0.608 **RHS** = 0.677

Instance 16: **LHS** = 0.258 + 0.133 + 0.24 = 0.631 **RHS** = 0.6645

This disparity between the whole weight calculated as sum of part weights and the actual Whole weight value shows us that this idea is NOT TRUE.



Viewer

Relation: abalone-weka.filters.unsupervised.attribute.Remove-R1-4,9

No.	1: Whole weight Numeric	2: Shucked weight Numeric	3: Viscera weight Numeric	4: **Shell weight** Numeric
1	0.514	0.2245	0.101	0.15
2	0.2255	0.0995	0.0485	0.07
3	0.677	0.2565	0.1415	0.21
4	0.516	0.2155	0.114	0.155
5	0.205	0.0895	0.0395	0.055
6	0.3515	0.141	0.0775	0.12
7	0.7775	0.237	0.1415	0.33
8	0.768	0.294	0.1495	0.26
9	0.5095	0.2165	0.1125	0.165
10	0.8945	0.3145	0.151	0.32
11	0.6065	0.194	0.1475	0.21
12	0.406	0.1675	0.081	0.135
13	0.5415	0.2175	0.095	0.19
14	0.6845	0.2725	0.171	0.205

Figure 1: A subset of the data with the weight attributes

The negative y-intercept value d = -0.0078, does not hold any meaning in this scenario, as the concept of negative weight is absurd. Finally,the value of the $CorrelationCoefficient = 0.9954$; indicates a strong positive relationship between the weight attributes.

3 Modelling Diameter - Weight Relationship

The two attributes of importance in this section are Diameter and Whole Weight. A scatter plot with Whole Weight as dependent variable and Diameter as independent variable is shown in Fig. 2. As is evident in the graph, the two attributes seem to share a non-linear relationship among each other. Let's fit the following models and choose the best one from them.

Let Diameter = D, and WholeWeight = W,

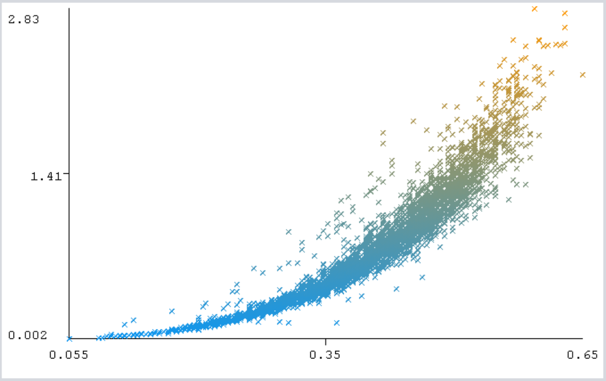


Figure 2: Scatterplot of the data: Diameter vs Whole Weight

a) Linear Model: $y = a * x + b$

A simple linear regression performed on the two attributes resulted in the following equation with a **Correlation Coefficient of** 0.9253.

$$W = 4.5731 * D - 1.0365 \tag{4}$$

b) Quadratic Model: $y = a * x + b * x^2 + c$

The **Correlation Coefficient** for the quadratic model was obtained as **0.9626**, which is a better value than in the linear model.

In Eq. (5),we can see that the slope for D is negative whereas the slope for D^2 term is positive. Refer Fig.4 for the plots of D^2 against Weight.

$$W = -3.3555 * D + 10.4968 * D^2 + 0.3477 \tag{5}$$

c) Cubic Model: $y = a * x^3$

The **Correlation Coefficient** for the quadratic model was obtained as **0.9904**, and the equation obtained is as below:

$$W = 10.3376 * D^3 \tag{6}$$

The most important thing to note for the Cubic relationship between Weight and Diameter is the high positive value of Correlation Coefficient. Such a value implies a strong correlation between the two features, and the highest among all the models, as we'll see later.

d) Exponential Model: $log(y) = a * x + b$

Finally, for the exponential model, we obtain a value of **0.9634** for the **Correlation Coefficient**.

The equation of the model:

$$logW = 8.1167 * D - 3.751 \tag{7}$$

Conclusion for Q3

Having obtained the Correlation coefficients of the models, and the graphs plotted for the same (refer Figures 3 - 6), we can see that the Cubic model is the best option to model the data and study the relationship between Diameter and Whole Weight.

In Figs. 3 - 6, the lines of fit for all the models are plotted on the Diameter vs Whole weight scatter plot. The model corresponding to the cubic model is one that fits the data the best.

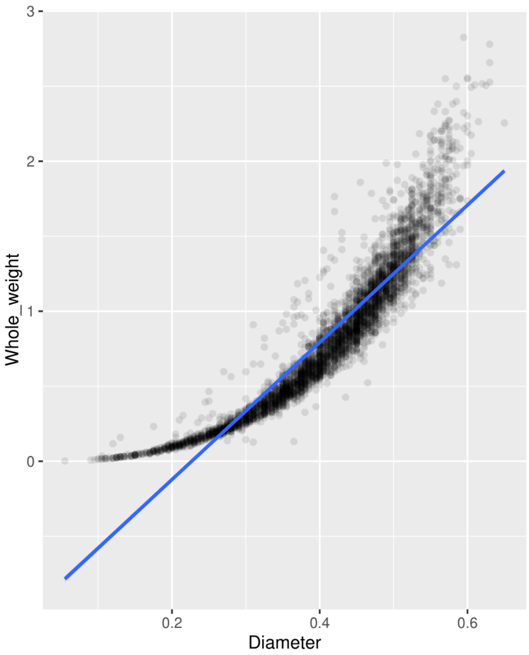


Figure 3: D vs. W

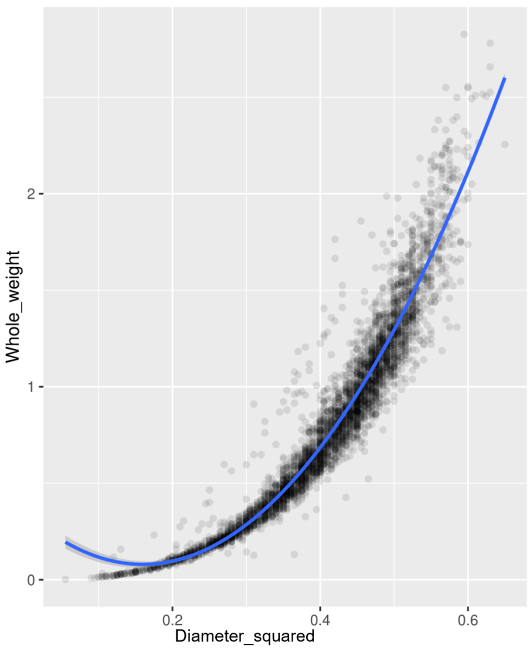


Figure 4: D^2 vs. W

Its correlation Coefficient is 0.9904, the highest among all models, and the graph plotted indicates a near perfect fit with respect to the scatter plot. Moreover, we can also understand this linearity between Whole Weight and Diameter_cubed as the Density measure (Rho) of the abalone, defined by the formula $\rho = Mass/Volume$, where mass is equivalent to Whole Weight, and Volume as a function of $Diameter^3$.

4 Logistic Regression

Preparing the dataset

The dataset has a Sex feature, nominal in nature, which takes 3 values- Male, Female and Infant. Since we do not use the Sex feature in our regression models below, the records of male and female sexes were combined to form the adult, and the new feature was renamed as AgeCategory. This attribute was defined in the abalone_updated.arff file as: $@attribute$ $'AgeGroup'$ $\{A, I\}$

While the sex category had approximately, equal number of records for each of its labels, the number of records in the adult category is more than double the number of records in infant category. Owing to the disproportionate representation of each category in the modified abalone dataset, we can expect a decreased performance of the regression model in predicting the AgeCategory of the specimen. The Sex feature did not have any missing values too.

a) Predicted from Length

The logistic regression model correctly classified 3271 instances resulting in an accuracy of **78.3098%**. Time taken to build model: 0.13 seconds.

Confusion Matrix :-

a	b	<– Classified as
2555	280	a = A
626	**716**	b = I

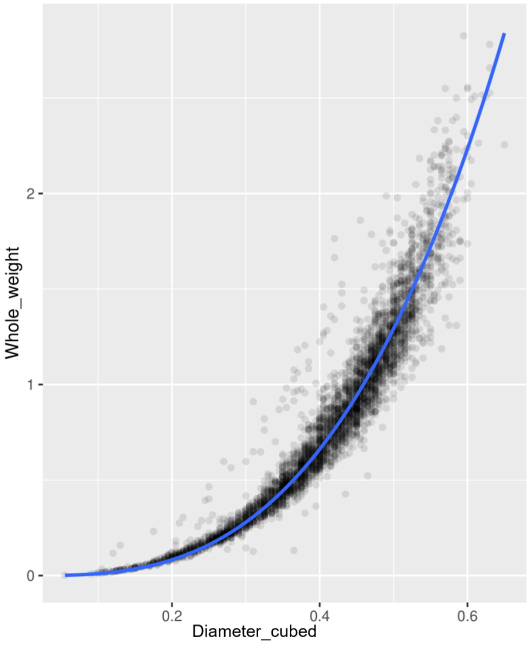


Figure 5: D^3 vs. W

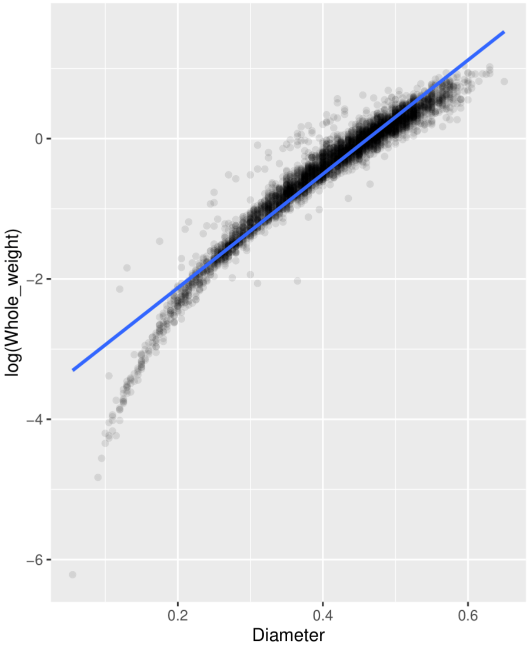
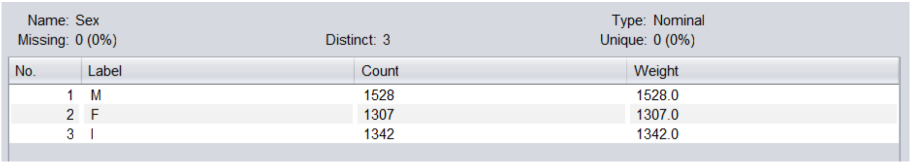


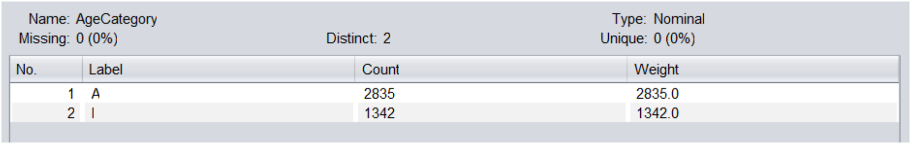
Figure 6: D vs. $log(W)$



Name: Sex
Missing: 0 (0%)
Distinct: 3
Type: Nominal
Unique: 0 (0%)

No.	Label	Count	Weight
1	M	1528	1528.0
2	F	1307	1307.0
3	I	1342	1342.0

Figure 7: Details about the Sex Feature



Name: AgeCategory
Missing: 0 (0%)
Distinct: 2
Type: Nominal
Unique: 0 (0%)

No.	Label	Count	Weight
1	A	2835	2835.0
2	I	1342	1342.0

Figure 8: Details about AgeCategory Feature

b) Age Category predicted from Whole Weight

The logistic regression model correctly classified 3325 instances resulting in an accuracy of **79.6026%**. Time taken to build model: 0.03 seconds.

Confusion Matrix :-

a	b	<– Classified as
2456	379	a = A
473	**869**	b = I

c) Age Category predicted from Class_Rings

The logistic regression model correctly classified 3292 instances resulting in an accuracy of **78.8125%** . Time taken to build model: 0.03 seconds

Confusion Matrix :-

a	b	<– Classified as
2644	191	a = A
694	**648**	b = I

d) Age Category predicted from Length, Whole Weight and Class Rings together:-

The logistic regression model correctly classified 3437 instances resulting in an accuracy of **82.2839%** . Time taken to build model: 0.11 seconds

Confusion Matrix :-

a	b	<– Classified as
2459	376	a = A
364	**978**	b = I

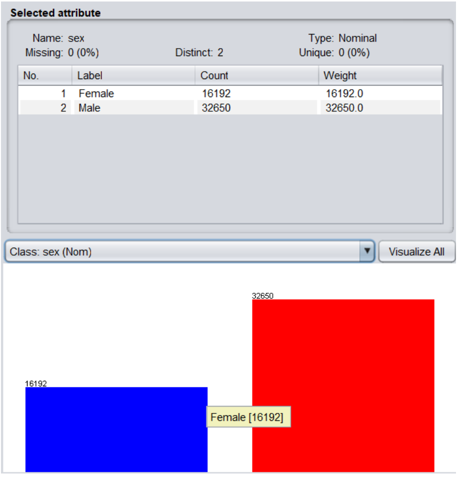
5 Adult Dataset

Adult is a multivariate dataset contains 48842 instances and 14 attributes. It is used for a Classification task to predict if a person earns over or below 50,000/- per year. This section studies the Sex feature in detail and prepares a simple and accurate model that can predict the sex of the person using the attributes available. The study of the Adult Dataset and the results of the coursework provided below were performed on Weka. **A ten-fold cross validation technique was adopted as the Test Mode**.

1. age
2. workingclass
3. fnlwgt
4. education
5. education-num
6. marital-status
7. occupation
8. relationship
9. race
10. sex
11. capital-gain

12. capital-loss
13. hours-per-week
14. native-country
15. Class (earning <50,000 or >50,000

The Adult.arff file is uploaded onto the WEKA Tool, and its basic features are studied. Since the feature to be predicted in this section is Sex, the corresponding legend for this feature is studied on the Preprocess tab. The Fig. 9, shows us that Blue and Red depict the Female and Male categories respectively. We can see that the records with Males in the dataset is a little more than double the number of Females in the dataset. This disparity in equal representation can affect the accuracy of the model too.


Selected attribute

Name: sex
Missing: 0 (0%)
Distinct: 2
Type: Nominal
Unique: 0 (0%)

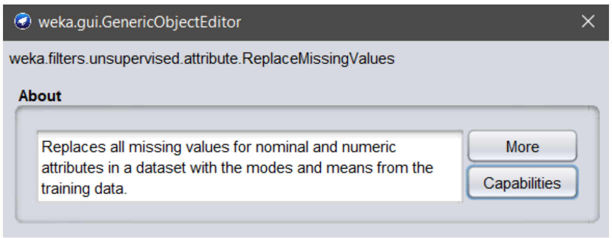
No.	Label	Count	Weight
1	Female	16192	16192.0
2	Male	32650	32650.0

Class: sex (Nom)

Figure 9: Sex Feature in Adult Dataset

Our aim is to find those attributes which enable us to predict the Sex feature with a good accuracy, meaning, we have to choose those attributes where there is a clear separation of male and female categories with respect to the labels of the attribute under consideration.

The following attributes- workingclass, occupation and native-country had varying levels of missing values. Weka Tool provides a filter than enables us to handle missing values in the dataset. Refer to Fig. 10.


weka.gui.GenericObjectEditor

weka.filters.unsupervised.attribute.ReplaceMissingValues

About

Replaces all missing values for nominal and numeric attributes in a dataset with the modes and means from the training data.

Figure 10: Inbuilt Filter to replace missing values

The baseline model considered consists of all attributes stated above to perform the prediction. The model performs with an accuracy of **84.501 %**, correctly classifying 41272 instances out of 48842 records. Before eliminating attributes from this baseline model in a brute-force manner, we can visualise and the the relationship between each attribute and Sex in the Preprocess tab on Weka Tool.

For example, in the case of **workingclass** attribute, both males and females appear in every label of the attribute. Such attributes do not provide a means to be used to predict the the Sex feature, because the Male and Females instances are not separable with respect to their Working Class label. Similar behaviour is observed from attributes-age, fnlwgt, education, education-num, capital loss and gain, marital status and Class. The following table contains the accuracy obtained when each of these attributes were removed from the model, separately, on by one.

Attribute Removed	Resulting Accuracy of model
age	84.5072 %
workingclass	84.2431 %
fnlwgt	84.2635 %
education	84.3987 %
education-num	84.501 %
marital status	84.2513 %
race	84.5625 %
capital-gain	84.5031 %
capital-loss	84.4969 %
native-country	84.4929 %
Class	84.3639 %

Accuracy of the Baseline Model = 84.501 %

The above attributes do not reduce the value of the accuracy of model by more than 1 %, therefore can be removed from the model as dead weight. Infact, removing two of these attributes, results in an increase in accuracy, implying that there is no need for them to be included into the model for predicting Sex. The following attributes can be retained to be further used for building the model:

- occupation
- relationship
- hours-per-week

Intuitively, we can see that all the above attributes involve gender roles in their definition; while occupation and hours-per-week mostly differ between men and women due to the predefined notions of which tasks are performed better by each of the sexes, relationship inherently revolves around the roles of men and women in society. It therefore tallies with the evidence we have obtained in favour of the mentioned attributes, and against the other attributes, which do not exhibit sufficient difference in behaviour between the sexes for prediction.

a) For the same reason, we can derive that attributes such as age, marital-status, capital-gain, capital-loss, education, education-num, working-class, fnlwgt, race and Class are the ones which can be eliminated without ever affecting the accuracy of the resulting model by more than 1 %.

b) The reason **relationship** assists the prediction of the Sex feature with the highest accuracy is the fact that it has two labels which correspond to Female and Male respectively to form a pure class(Refer to Fig: 11). In the Adult dataset, the label Wife(FEMALE) in relationship has a count of 2331, and the label Husband(MALE) has a count of 19716. This clear demarcation of the Sex feature among the labels of relationship attribute gives rise to an impressive accuracy of prediction.

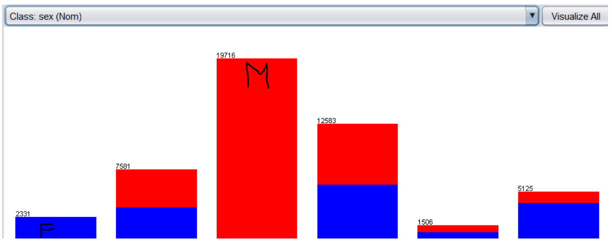

Class: sex (Nom)

Figure 11: Relationship vs Sex

The following table shows the confusion matrix for the baseline model, which predicts with an accuracy of 84.501 %. Comparing this matrix to the one we get with only these three attributes- Relationship, Occupation, and Hours/week, we can see that there more true positives for Females in our new model even though its accuracy is slightly lesser- 83.6227 %. In other words, there are more records with Sex attribute correctly classified as Females. But corroborating with the dip in accuracy value, we see a decrease in the number of true positives for male category.

Confusion Matrix for baseline model:-

a	b	<– Classified as
13033	3159	a = Female
4411	**28239**	b = Male

The accuracy of the model after removing the above attributes is **83.6227 %**.

Confusion Matrix for improved model :-

a	b	<– Classified as
13218	2974	a = Female
5025	**27625**	b = Male

c) In specific, for native-country = Holland, the weight is so high, because only one record has the country as Holand, with a value of Female for the sex attribute. This misguides the model to give this nominal category a higher weight, as it will be able to predict the sex of adult from Holand as female.

While it is true that using all attributes contributes to a higher accuracy, it can be argued that building a model with as much as 14 attributes, for an increase of only 1% in accuracy is not justified, as compared to just two attributes. If the dependency of the Sex feature on two attributes is sufficient to get an accuracy of 83.6227%, then the goal of building both a simple and accurate model is achieved.

Attributes	Accuracy of model
All Attributes	84.501%
relationship, **occupation**, **hours-per-week**	83.6227%

6 Appendix

6.1 Answer 4:

a) Age Category Predicted from Length

Detailed Accuracy by Class :-

Class	TP Rate	FP Rate	Precision	Recall	F-Measure
Adult (A)	0.901	0.466	0.803	0.901	0.849
Infant (I)	0.534	0.099	0.719	0.534	0.612
Weighted Avg.	0.783	0.348	0.776	0.783	0.773

Other Performance Metrics

Incorrectly Classified Instances	906 (21.6902 %)
Kappa statistic	0.4664
Mean absolute error	0.2996
Root mean squared error	0.3882
Relative absolute error	68.6873 %
Root relative squared error	83.1287 %

b) Age Category predicted from Class_Rings

Detailed Accuracy by Class :-

Class	TP Rate	FP Rate	Precision	Recall	F-Measure
Adult (A)	0.866	0.352	0.839	0.866	0.852
Infant (I)	0.648	0.134	0.696	0.648	0.671
Weighted Average	0.796	0.282	0.793	0.796	0.794

Other Performance Metrics

Incorrectly Classified Instances	852 (20.3974 %)
Kappa statistic	0.5235
Mean absolute error	0.2754
Root mean squared error	0.3715
Relative absolute error	63.1447 %
Root relative squared error	79.5465 %

c) Age Category predicted from Class_Rings

Detailed Accuracy by Class :-

Class	TP Rate	FP Rate	Precision	Recall	F-Measure
Adult (A)	0.933	0.517	0.792	0.933	0.857
Infant (I)	0.483	0.067	0.772	0.483	0.594
Weighted Average	0.788	0.373	0.786	0.788	0.772

Other Performance Metrics

Incorrectly Classified Instances	885 (21.1875 %)
Kappa statistic	0.461
Mean absolute error	0.3231
Root mean squared error	0.3957
Relative absolute error	74.0867 %
Root relative squared error	84.7446 %

d) Age Category predicted from Length, Whole Weight and Class Rings together:-

Detailed Accuracy by Class :-

Class	TP Rate	FP Rate	Precision	Recall	F-Measure
Adult (A)	0.867	0.271	0.871	0.867	0.869
Infant (I)	0.729	0.133	0.722	0.729	0.726
Weighted Average	0.823	0.227	0.823	0.823	0.823

Other Performance Metrics

Incorrectly Classified Instances	740 (82.2839 %)
Kappa statistic	0.5947
Mean absolute error	0.2571
Root mean squared error	0.3553
Relative absolute error	58.9495 %
Root relative squared error	76.0962 %