ABALONE AGE PREDICTION

PROBLEM STATEMENT:-

ABALONES ARE ENDANGERED MARINE SHELLS THAT ARE FOUND IN THE COLD COASTAL WATER WORLD WIDE, BEING DISTRIBUTED OFF THE COASTS OF NEW ZEALAND, SOUTH AFRICA, AUSTRALIA, WESTERN NORTH AMERICA, AND JAPAN. THEY ARE HIGHLY NUTRITIOUS , ELEGANCE, RARE FOOD WHICH IS GOOD TO EAT IS EXTENSIVELY CONSUMED IN FRANCE, NEW ZEALAND, CERTAIN PARTS OF LATIN AMERICA, JAPAN AND KOREA. THE SHELLS OF ABALONE ARE USED FOR DECORATIVE PURPOSE. THEREFORE , ABALONE IS ECONOMICALLY SIGNIFICANT. THE PRICE OF ABALONE IS POSITIVELY CORRELATED TO ITS AGE. HOWEVER DETERMINING THE AGE OF ABALONE IS VERY DIFFICULT PROCESS. RINGS ARE FORMED IN THE INNER SHELL OF THE ABALONE. ONE RING WILL BE DEVELOPED PER YEAR. THE AGE OF ABALONE IS DETERMINED BY CUTTING THE SHELL THROUGH THE CONE AND STAINING IT AND COUNTING THE NUMBER OF RINGS THROUGH A MICROSCOPE.

DATA ANALYSIS:-

THE ABALONE DATASET IS A DATASET THAT CONTAINS MEASUREMENTS OF PHYSICAL CHARACTERISTICS OF DIFFERENT ABALONES. IT HAS 4177 INSTANCES.

IN THIS SECTION THE DISTRIBUTION OF EACH ATTRIBUTE IS ANALYZED INDIVIDUALLY. WE START ANALYZING THE DISTRIBUTION OF THE TARGET ATTRIBUTE RINGS. THE REST OF THE ATTRIBUTES ARE DIVIDED IN GROUPS FOR CONVENIENCE OF THE ANALYSIS: A GROUP CALLED SIZE, CONTAINING ATTRIBUTES THAT REPRESENTS THE DIMENSIONS OF AN ABALONE, A GROUP WEIGHT, CONTAINING THE DIFFERENT WEIGHT ATTRIBUTES AND A THIRD GROUP COMPOSED ONLY OF THE SEX ATTRIBUTE. THE CONTINOUS OR QUANTITATIVE ATTRIBUTES WERE ANALYZED USING HISTOGRAMS AND BOXPLOTS, WHILE CATEGORICAL ATTRIBUTES WERE ANALYZED USING BARPLOTS.

THE TARGET ATTRIBUTE

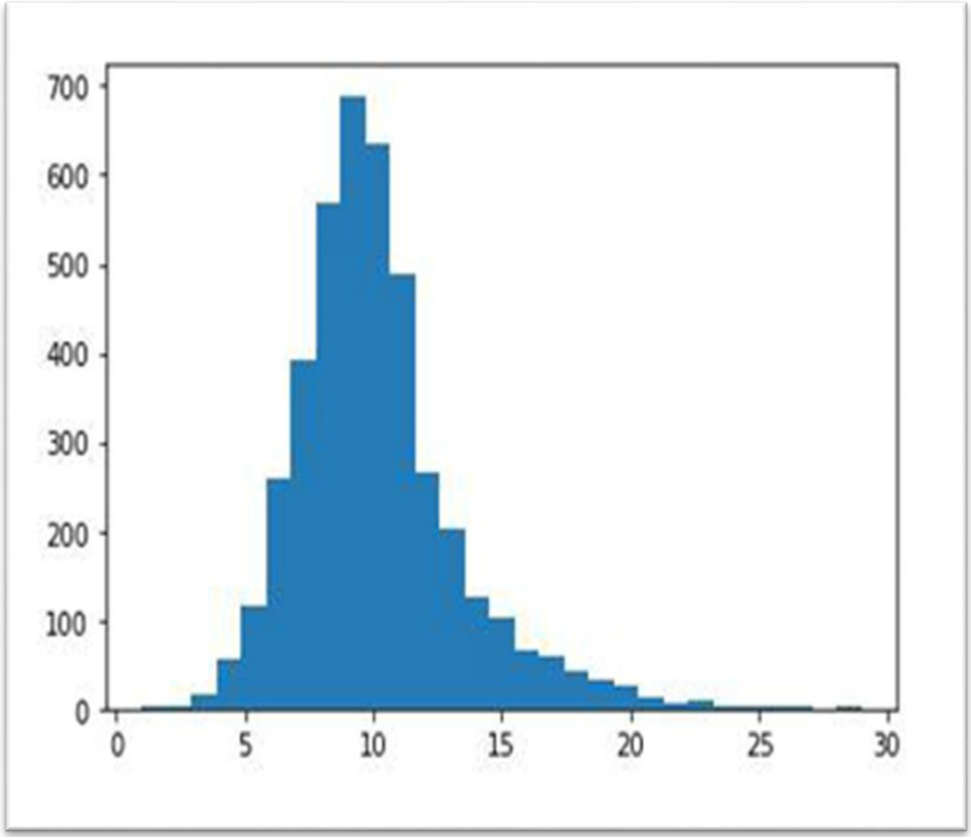
THE ANALYSIS SHOWS THAT THE RING ATTRIBUTE VALUES RANGES FROM 1 TO 29 RINGS ON AN ABALONE SPECIMEN. HOWEVER, THE MOST FREQUENT VALUES OF RINGS ARE HIGHLY CONCENTRATED AROUND THE MEDIAN OF THE DISTRIBUTION, SO THAT, THE 2ND AND 3RD QUARTILES ARE DEFINED IN A RANGE OF LESS THAN 1

STD DEVIATION. WE OBSERVE THAT ITS POSSIBLE TO APPROXIMATE THE

DISTRIBUTION OF THIS ATTRIBUTE TO A NORMAL CURVE.

THE DISTRIBUTION OF RINGS IN ABALONE DATASET IS SHOWN IN FIGURE

DISTRIBUTION OF RINGS



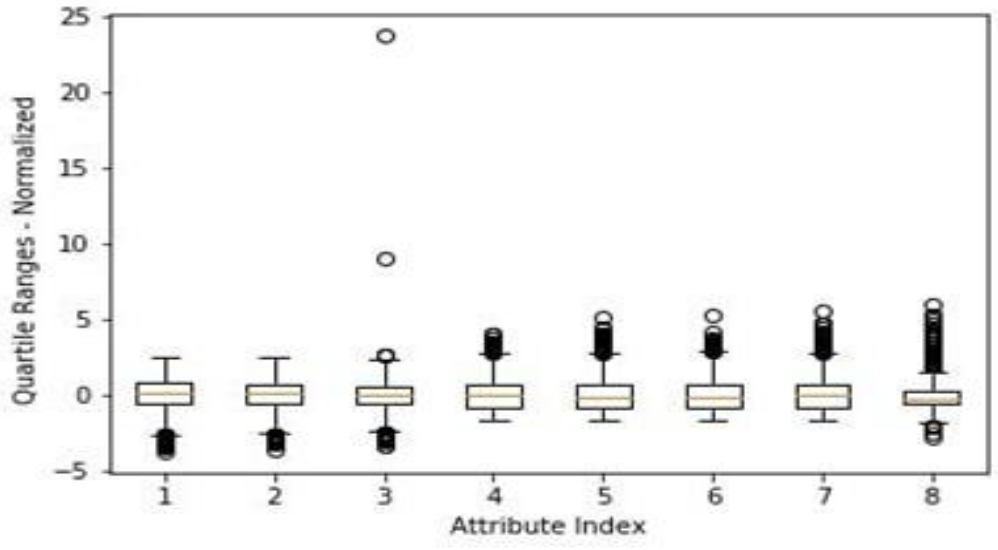
BOXPLOT:-

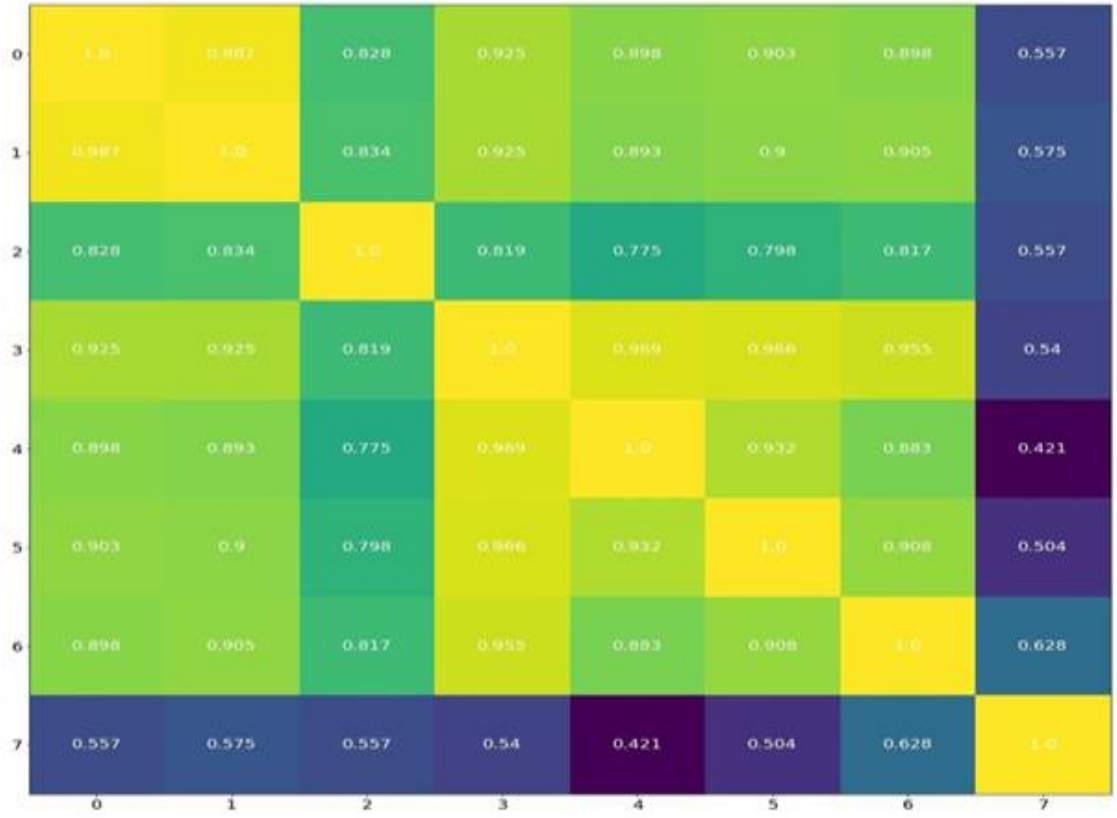
THE MINIMUM, MAXIMUM, MEAN, MEDIAN, STANDARD DEVIATION AND

INTERQUARTILE RANGE OF ALL THE NUMERIC ATTRIBUTES ALONG WITH

DEPENDENT VARIABLE OF THE DATASET IS CALCULATED AND PLOTTED USING A

BOXPLOT FOR EASY VISUALIZATION OF OUTLIERS. DUE TO THE LARGER RANGE OF "RINGS" VARIABLE, AN UN NORMALIZED BOXPLOT RENDERS THE OTHER VARIABLES' BOXPLOTS INCOMPREHENSIBLE BY SQUEEZING THEIR RANGES. TO BRING ALL THE VARIABLES ON THE SAME SCALE, THEY ARE NORMALIZED SUCH THAT THEY ALL HAVE ZERO MEAN AND STANDARD DEVIATION 1. . THE ATTRIBUTES LENGTH AND DIAMETER HAVE ALMOST THE SAME NORMALIZED RANGE WHILE THERE ARE A FEW OUTLYING VALUES FOR THE HEIGHT ATTRIBUTE WHICH MIGHT MAKE THE TASK OF REGRESSION DIFFICULT. ALL THE WEIGHT ATTRIBUTES ALSO HAVE ALMOST THE SAME NORMALIZED RANGE. THE RINGS LABEL IS NOT ANALYZED SINCE IT WILL BE USED IN AN UN NORMALIZED FORM FOR THE REGRESSION TO OBTAIN A PR





HEAT MAP

METHODOLOGY:

THE FIRST STEP TOWARDS APPLYING LINEAR REGRESSION TO PREDICT THE AGE OF ABALONE WAS TO NUMERICALLY CODE THE SEX VARIABLE IN THE DATASET. FOR THIS, THE FOLLOWING APPROACH WAS TAKEN:

1.TWO ATTRIBUTES WERE CREATED IN PLACE OF SEX. LET THEM BE S1 AND S2.

2.THE SEX VALUE OF EACH SAMPLE WAS CHECKED

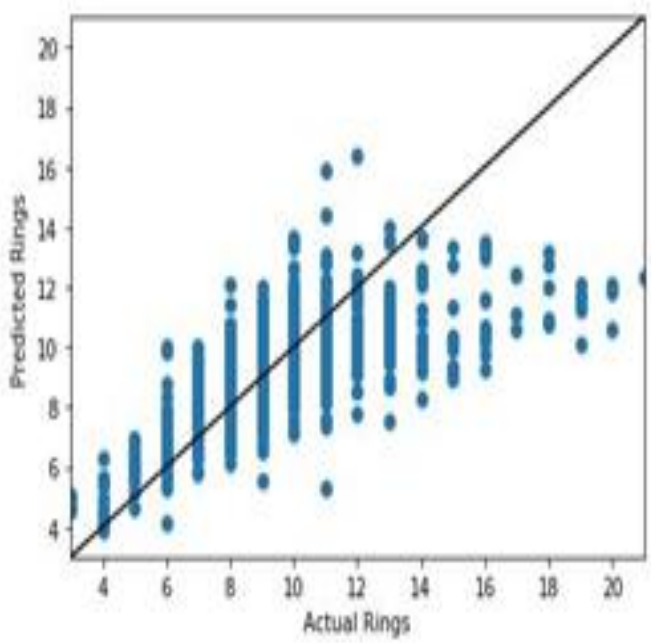
3.IF IT IS A MALE, THEN S1 IS EQUATED TO 1 AND S2 IS EQUATED TO 0.

4. IF IT IS FEMALE, THEN S1 IS EQUATED TO 0 AND S2 IS EQUATED TO 1.

5.IF IT IS INDETERMINATE, BOTH ARE KEPT 0

RESULTS AND CONCLUSIONS:-

THE MEAN ABSOLUTE ERROR FOR THE VARIOUS MODELS WITH THE CORRESPONDING HYPER-PARAMETERS IN BRACKETS. THE RESULTS FOR 10 FOLD CV HAS BEEN AVERAGED ON THE TEN FOLDS. OTHER PENALIZED REGRESSION METHODS WHICH WERE TIRED WERE LASSO LAR AND ELASTIC NET. ALL OF THESE PERFORMED SIMILAR TO RIDGE REGRESSION WITH THEIR BEST PERFORMANCE BEING BIT LOWER THAN THE PERFORMANCE OF OLS. IT IS CLEARLY VISIBLE THAT RANSAC YIELDED THE PERFORMANCE.



OLS CV BEST MODEL

OLS:-

OLS (WITHOUT SMOTHE):-1.500

OLS(WITH SMOTHE):-1.875

OLS (WITH 10 FOLD CV):-1.636

RIDGE:-

RIDGE(WITHOUT SMOTHE):-1.499(ALPHA=0.1)

RIDGE(WITH SMOTHE):-1.882(ALPHA=0.01)

RIDGE(WITH 10 FOLD CV):-1.627(ALPHA=0.01)