Q1. Given the data set below, compute descriptive statistics of the score (mean, median, std deviation, as well as the 95% confidence interval. Show how you perform your computations (no SPSS Modeler), and then verify them using Modeler.

Exam	Score
1	84
2	81
3	77
4	90
5	92
6	64
7	75
8	34
9	75
10	60

Computing Mean: Mean is the sum of the values divided by the number of values

Mean = ∑(Score)/No. of exams = 732/10 = 73.2

Computing Median: it is the midpoint of the scores, where the scores above and below median are equal when they are put in an order either in ascending or in descending order

Values in ascending order: 34, 60, 64, 75, **75, 77**, 81, 84, 90, 92

Median = (75+77)/2 = **76**

Computing standard deviation: It is used to measure the variability or spread of the score around mean

$$s = \sqrt{\frac{\sum (x-\bar{x})^2}{n-1}}$$

= SQRT (((34-73.2)^2+(60-73.2)^2+(64-73.2)^2+(75-73.2)^2+(75-73.2)^2+(77-73.2)^2+(81-73.2)^2+(84-73.2)^2+(90-73.2)^2+(92-73.2)^2)/9)

= **17.09**

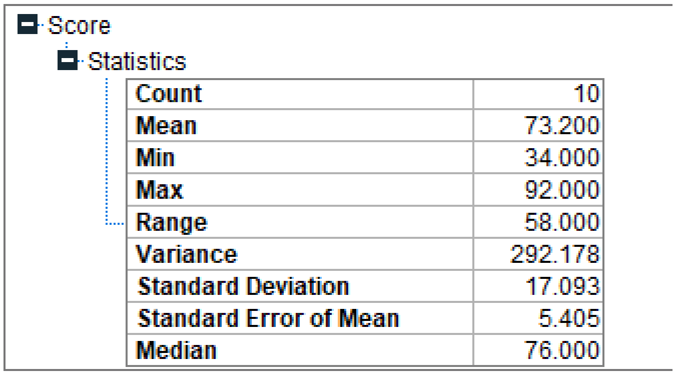
Computing 95% confidence interval: It is an estimate where 95% of the values are going to lie around the mean. Since the size is small with n=10, t statistic should be used instead of Z score while computing the confidence interval.

$\bar{x} \pm t_{\alpha/2}(s/\sqrt{n})$ where α = 1-0.95 = 0.05 also, t distribution depends of degrees of freedom df = 10-1 = 9

= 73.2 ± 2.262(17.09/sqrt(10)) = 73.2 ±12.22

Implying that with 95% of confidence around the mean score can go between **84.42 and 60.98** with margin error of 12.22 (Standard error of mean = $s/\sqrt{n}$ = 5.404)

When checked with SPSS modeler, the below are the values for mean, median, standard deviation and standard error of mean. Margin of error = $t_{\alpha/2}$ * Standard error of mean;



Count	10
Mean	73.200
Min	34.000
Max	92.000
Range	58.000
Variance	292.178
Standard Deviation	17.093
Standard Error of Mean	5.405
Median	76.000

Hence, the calculated and SPSS modeler values are same.

Q2. Given the following dataset of MBA students, build a KNN model to classify two new applicants (Diego Maradona, Lionel Messi). Use K=3, and Manhattan distance. DO it manually (no SPSS Modeler) You can find the data in MBA.xlsx

After clear observation of MBA data, since there is only rating 1, so the chance of classifying the new applicants is high on rating 1. But Let me give a manual description of it

Student	Rating	GPA	GMAT
1	1	2.96	671
2	1	3.14	548
3	1	3.22	557
4	1	3.29	602
5	1	3.69	580
6	1	3.46	768
7	1	3.03	701
8	1	3.19	738
9	1	3.63	522
10	1	3.59	663

NAME	GPA	GMAT
Diego Maradona	3.02	450
Lionel Messi	3.95	551

Since the values of GPA and GMAT are in varied ranges, considering min-max standardization, min(GPA)= 2.96 and Max(GPA) = 3.69; min(GMAT) =450 and Max(GMAT) =768

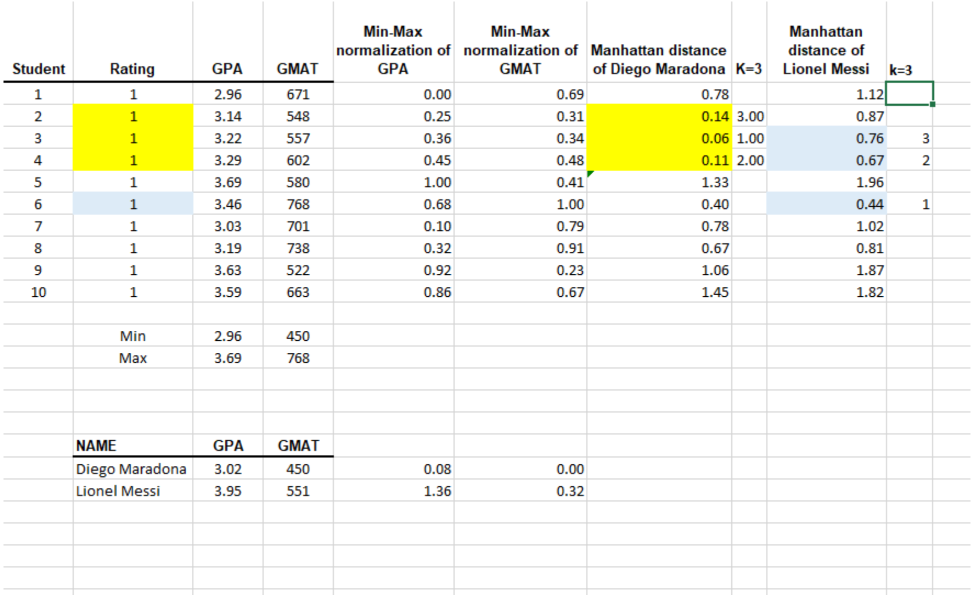
From the above the normalization is done by

$$\text{Min - Max Normalization} = \frac{X - \min(X)}{\max(X) - \min(X)}$$

And the Manhattan distances are calculated from each applicant to the student in the attached MBA sheet

Formula for calculating Manhattan distance is

$$d(i,j)=|x_{1i}-x_{1j}|+|x_{2i}-x_{2j}|$$



Student	Rating	GPA	GMAT	Min-Max normalization of GPA	Min-Max normalization of GMAT	Manhattan distance of Diego Maradona	K=3	Manhattan distance of Lionel Messi	k=3
1	1	2.96	671	0.00	0.69	0.78		1.12	
2	1	3.14	548	0.25	0.31	0.14	3.00	0.87	
3	1	3.22	557	0.36	0.34	0.06	1.00	0.76	3
4	1	3.29	602	0.45	0.48	0.11	2.00	0.67	2
5	1	3.69	580	1.00	0.41	1.33		1.96	
6	1	3.46	768	0.68	1.00	0.40		0.44	1
7	1	3.03	701	0.10	0.79	0.78		1.02	
8	1	3.19	738	0.32	0.91	0.67		0.81	
9	1	3.63	522	0.92	0.23	1.06		1.87	
10	1	3.59	663	0.86	0.67	1.45		1.82	
	Min	2.96	450						
	Max	3.69	768						
	NAME	**GPA**	**GMAT**						
	Diego Maradona	3.02	450	0.08	0.00				
	Lionel Messi	3.95	551	1.36	0.32				

The distances are ordered in ascending order, for k=3, the nearest neighbors for the two applicants are marked in the above. Considering the rating factor, since the majority (actually all) are with rating 1, there is highest chance for new applicants be rated as 1.

Q3. A charitable organization wants to create a predictive model of the donations it receives, and for that purpose it collects data on donors' education (in years), annual income (in $1,000) and the number of children the donor has. The charity developed a multiple linear regression model with these three variables. The table below summarizes these results

Variables	F_Statistic	R^2	Adjusted -R^2	S_e	Parameter Estimates, p-value
Education X1),	7.974	0.825119	0.813714	61.9002	b_0 = 445.0583, p<0.001
Income (X2),	(p< 0.001)	b_1 = 31.79566, p<0.001 and			b_2 = 0.003698, p<0.001
Kids (X3)					b_3 = 45.69131, p<0.001

Se = standard error of the estimate

a) Write and explain the regression equation.
b) Explain the performance metrics of the regression equation using the table above.
c) Predict the expected donation by a person with 16 years of education, $90,000 annual income and 2 children

a) Regression equation is written in the form y= $b_0 + b_1x_1 + b_2x_2 + b_3x_3$

=>donations = 445.0583 + 31.79566*Education + 0.003698*Income + 45.69131*Kids

Intercept 445.0583 holds a positive relationship on donations with education, income and kids.

Unit increase in period of education results in increase of 31.79566 units of donations, when income and children of donors are kept constant; Unit increase in annual income of donors results in increase of 0.003698 donations, when education of donors and count of their children are kept constant; Unit increase in number of children of donors results in increase of 45.69131 donations, when income and education criteria of donors are kept constant.

b) From the table it is seen that R^2 is 82.51%, which represents proportion of variability in response to predictive set, donations and donors. That is 82.51% of variability in donations is accounted by linear relation with donors education period, income and children.

It is seen that adjusted R^2<R^2, indicating one predictor is irrelevant and can be omitted. (which would be income as it holds least slope when compared to other two)

Standard error of mean from table is 61.90 units, which is an estimate error in donation based on the regression dependents. And this error can be decreased when a useful predictor is added to the model.

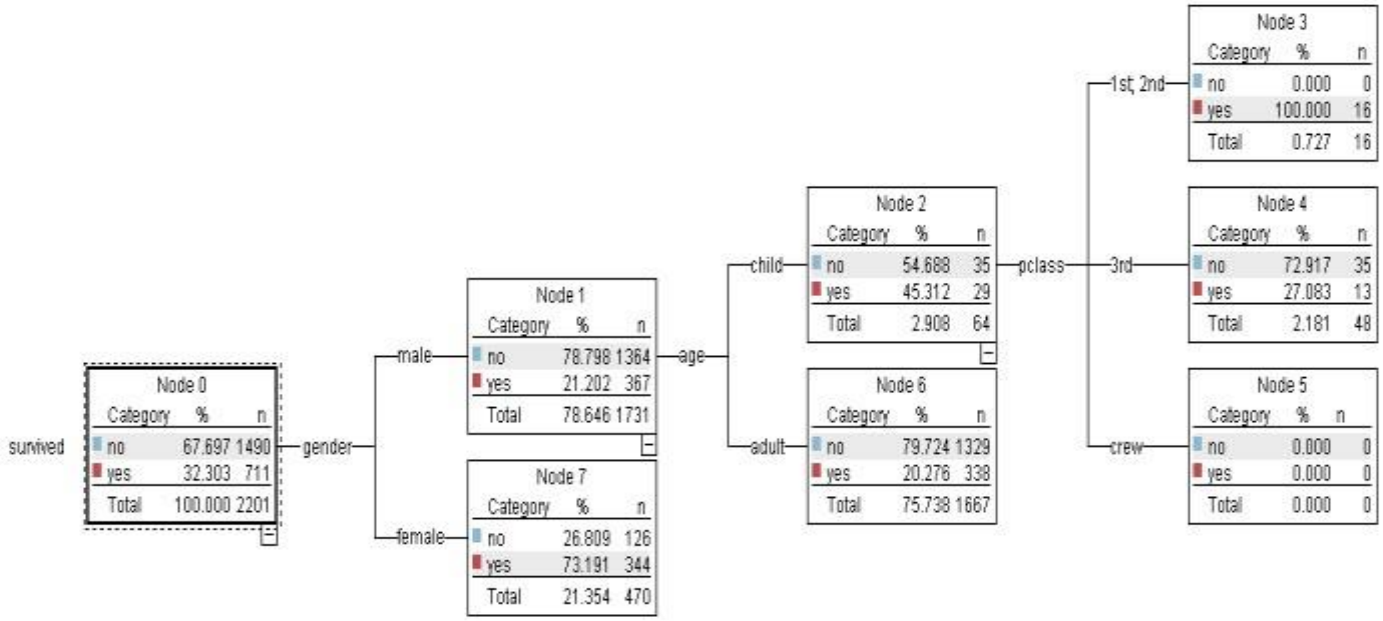
F is 7.974 with no significant p value, shows that variables in the model are significant.

c) the regression equation is donations = 445.0583 + 31.79566*Education + 0.003698*Income + 45.69131*Kids = 445.0583 + 31.79566*16 + 0.003698*90 + 45.69131*2 = 1045.5043 units

Q4. The **titanic dataset** describes the survival status of individual passengers on the Titanic. The variables on the dataset are all nominal: pclass, survived, age, and gender.

✦ pclass refers to passenger class (1st, 2nd, 3rd, crew), and is a proxy for socio-economic class. ✦ age is dichotomized at adult vs. child.
✦ gender is male or female. ✦ survived is yes or no
The dataset has 2201 instances with no missing data.

A C5.0 algorithm is applied on the whole data set, yielding the decision tree below:



Node 0

Category	%	n
🟦 no	67.697	1490
🟥 yes	32.303	711
Total	100.000	2201

Node 1

Category	%	n
🟦 no	78.798	1364
🟥 yes	21.202	367
Total	78.646	1731

Node 7

Category	%	n
🟦 no	26.809	126
🟥 yes	73.191	344
Total	21.354	470

Node 2

Category	%	n
🟦 no	54.688	35
🟥 yes	45.312	29
Total	2.908	64

Node 6

Category	%	n
🟦 no	79.724	1329
🟥 yes	20.276	338
Total	75.738	1667

Node 3

Category	%	n
🟦 no	0.000	0
🟥 yes	100.000	16
Total	0.727	16

Node 4

Category	%	n
🟦 no	72.917	35
🟥 yes	27.083	13
Total	2.181	48

Node 5

Category	%	n
🟦 no	0.000	0
🟥 yes	0.000	0
Total	0.000	0

Write the rules that can be derived from the decision tree together with the support and the confidence of each rule

The output decision tree shows that starting from a decision node of survivals, with support of 100% data and leaving no data. 711 out of 2201 has survived, which derives to confidence of 32.303%. when the tree is grown by placing the gender under survivals, under males, it is seen that the confidence is 21.202% with 367 passengers surviving out of 1731 with support of 367/2201 = 16.67% which is 367 male survivors out of 2201. Where as 78.646% is the support for Male node. Similarly, for female node, out of 470 records, with 73.191% confidence, 344 female out of 470 has survived with support of 344/2201 = 15.63% where as 21.354% is the percentage of support of full female node.

The following table summarizes the decision rules

Antecedent	Consequent	Support	Confidence
If(gender=male) ^ (age=child) ^ (pclass=1st,2nd)	Then survived=Yes	16/2201	16/16
If(gender=male) ^ (age=child) ^ (pclass=3rd)	Then Survived= No	35/2201	35/48
If(gender=male) ^ (age=child) ^ (pclass=Crew)	Then survived = No	0/2201	0/0

If(gender=male) ^ (age=adult)	Then survived = No	1329/2201	1329/1667
If(gender=female)	Then Survived = Yes	344/2201	344/470

For Child node, with support of 2.908%, there are 64 children out of 2201.

Q5. The dataset in the Excel file consists 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.

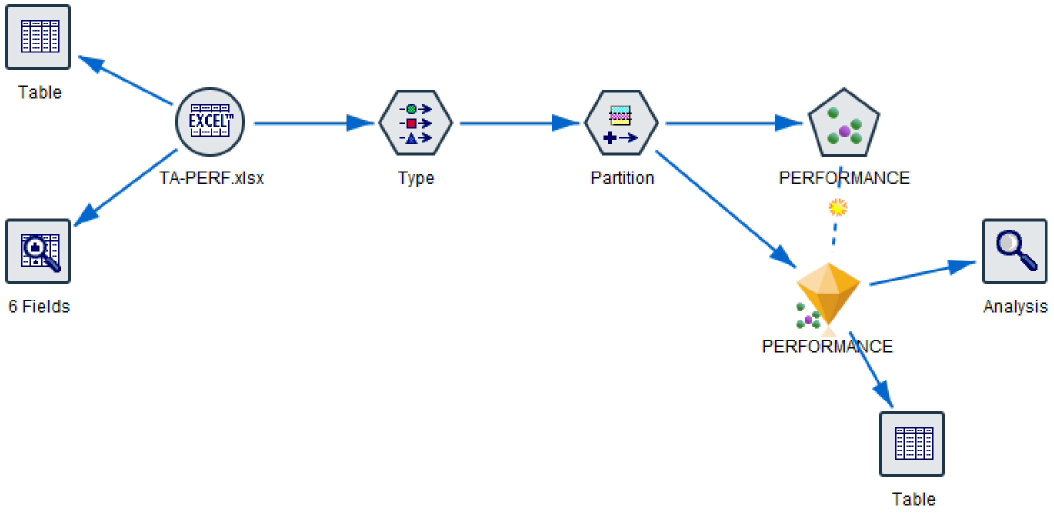
Attribute Information:

1. Whether or 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 (Source: UCI Machine Learning Repository)

Build a KNN classification model using the attached dataset. Report the predictive accuracy of the model.

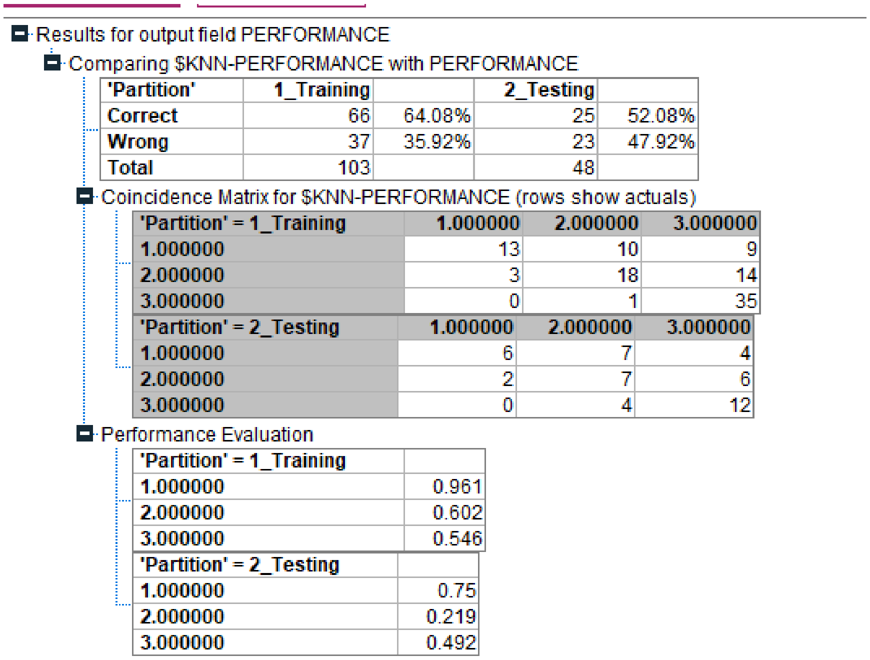
(Use SPSS Modeler to complete this question)

The following model is executed in the modeler for running a KNN classification model, attached as answer_5.str



The TA-PERF sheet is read for nominal values of performance (Class) attribute and binary values of semester Eng-Spkr attribute. The model is partitioned to 70% of training and 30% of testing and a KNN node is added and run with performance as target on all other attributes, taking k from 3 to 5.

When analysis node is attached to the diamond node generated as a result of KNN distribution, the following seen



Results for output field PERFORMANCE

Comparing $KNN-PERFORMANCE with PERFORMANCE

'Partition'	1_Training		2_Testing	
Correct	66	64.08%	25	52.08%
Wrong	37	35.92%	23	47.92%
Total	103		48	

Coincidence Matrix for $KNN-PERFORMANCE (rows show actuals)

'Partition' = 1_Training	1.000000	2.000000	3.000000
1.000000	13	10	9
2.000000	3	18	14
3.000000	0	1	35

'Partition' = 2_Testing	1.000000	2.000000	3.000000
1.000000	6	7	4
2.000000	2	7	6
3.000000	0	4	12

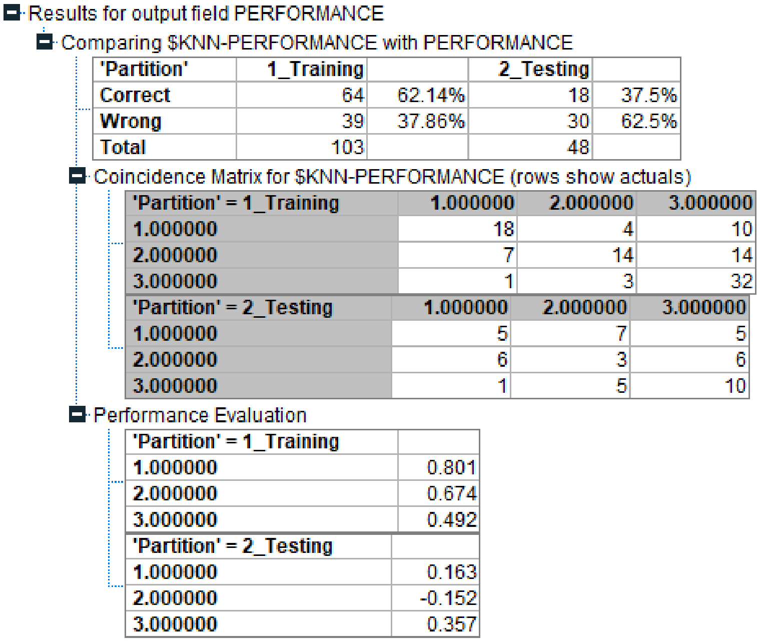
Performance Evaluation

'Partition' = 1_Training	
1.000000	0.961
2.000000	0.602
3.000000	0.546

'Partition' = 2_Testing	
1.000000	0.75
2.000000	0.219
3.000000	0.492

Since there is no unbalanced data seen in the matrix, accuracy can be considered as an performance metric. The accuracy for the test data is (6+7+12)/(6+7+4+2+7+6+4+12) = 52.08% which is very low training data accuracy, 64.08%, determining the chance of overfitting.

When I increased k value, within 10 and 15, though error rate decreased, accuracy on testing was around 35% compared to training data of 60% approximately.



Results for output field PERFORMANCE

Comparing $KNN-PERFORMANCE with PERFORMANCE

'Partition'	1_Training		2_Testing	
Correct	64	62.14%	18	37.5%
Wrong	39	37.86%	30	62.5%
Total	103		48	

Coincidence Matrix for $KNN-PERFORMANCE (rows show actuals)

'Partition' = 1_Training	1.000000	2.000000	3.000000
1.000000	18	4	10
2.000000	7	14	14
3.000000	1	3	32

'Partition' = 2_Testing	1.000000	2.000000	3.000000
1.000000	5	7	5
2.000000	6	3	6
3.000000	1	5	10

Performance Evaluation

'Partition' = 1_Training	
1.000000	0.801
2.000000	0.674
3.000000	0.492

'Partition' = 2_Testing	
1.000000	0.163
2.000000	-0.152
3.000000	0.357