**Data Mining & Predictive Analytics**

**Midterm Exam**

# Spring 2022

*INSTRUCTIONS*: Please answer **four** questions of your choice in the exam (25 points each) The completed exam must be returned in no more than 24 hrs. (Note: although the problem statements are long, the questions are straight forward. The exam can be completed in a few hours. So you should not have much problem with it)

Completed exams must be posted in your student’s drop box in the ilearn website in a zip file labeled *studentname\_midtermexam.*

In preparing your answers, you may use and reference published or unpublished sources, the World Wide Web, and/or your own notes. However, you may not ask for or receive help from any person on the exam. Write-ups should be delivered in Word, PDF and/or Excel format.

When you return the exam, you will be required to include a cover sheet (look in the next page) with a statement indicating that the exam is entirely your own work. Your name should appear on this cover sheet

A comment on academic honesty (may sound obsessive, but it is good advice):

As I have mentioned in previous occasions, the nature of online work sometimes leads to questionable decision making. This is an exam, and therefore it is individual work. **Please avoid the temptation of working in groups**. It is not only morally wrong, it is also myopic, as group work presented as individual work stands out, no matter how much you try to disguise it.

**Before doing something silly, please read Marist College Academic Integrity Policy.** <https://www.marist.edu/academic-resources/advising/academic-integrity-policy>Good luck!!

EL //

**NOTE: As I usually do in my courses, I have posted a** [**short anonymous survey**](https://docs.google.com/forms/d/e/1FAIpQLSdpiv-dRWRtRyB9NWVJyqtSrvAF3Z8JdM2P5Nf-cg9oZSfHBw/viewform?usp=sf_link)  **that I would like you to fill out. I use it to have a better feeling of how the course has gone and to get some feedback and ideas from you. So please, if you can take a few minutes to complete it, I would really appreciate it.**

**Data Mining & Predictive Analytics**

# Midterm Exam Spring 2022

I hereby certify that I have completed the attached examination materials, using only my own efforts. I have not asked for or received help from any person in completing this exam.

Swetha Adike 3/3/2022

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_

(Name) (date)

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

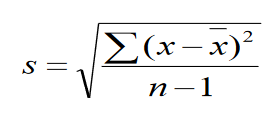
Mean = ∑(Score)/No. of exams = 732/10 = 73.2

**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



= 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.

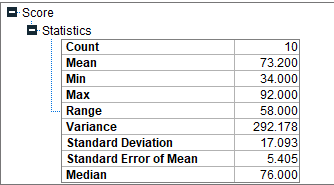
 where α = 1-0.95 = 0.05 also, t distribution depends of degrees of freedom

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/ = 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 = tα/2 \* Standard error of mean;



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

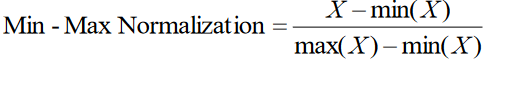
(Diego Maradona, Lionel Messi). Use K=3, and Manhattan distance. DO it manually (no SPSS Modeler) You can find the data in MBA.xlsx

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 | **NAME** | **GPA** | **GMAT** |
| 6 | 1 | 3.46 | 768 | Diego Maradona | 3.02 | 450 |
| 7 | 1 | 3.03 | 701 | Lionel Messi | 3.95 | 551 |
| 8 | 1 | 3.19 | 738 |
| 9 | 1 | 3.63 | 522 |
| 10 | 1 | 3.59 | 663 |

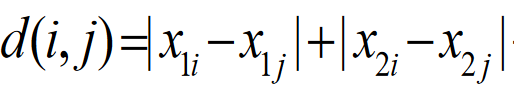
Since the values of GPA and GMAT are in varied ranges, considering min-max standardization, min(GPA)= 2.96 and Max(GPA) = 3.69; min(GMAT) =450 and Max(GMAT) =768

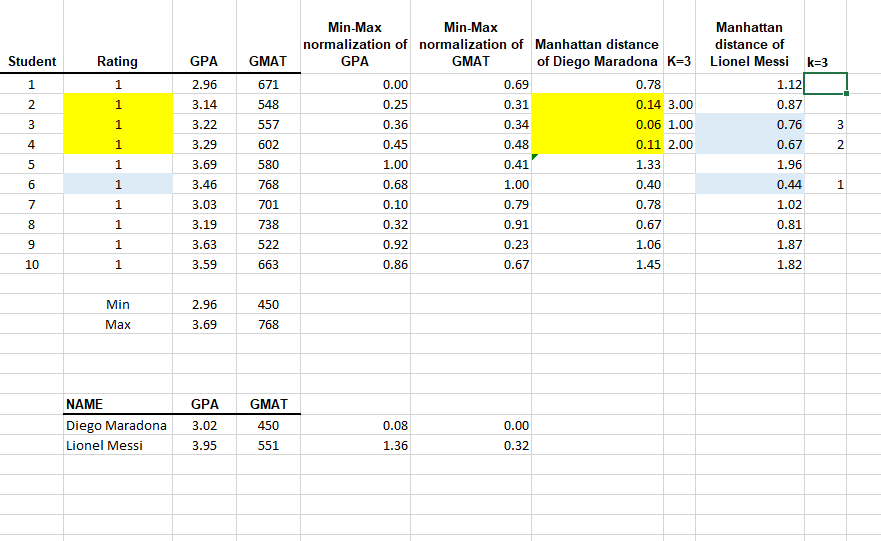
From the above the normalization is done by



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

Formula for calculating Manhattan distance is





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

Adjusted

Variables F\_Statistic R2 -R2 Se Parameter Estimates, p-value

Education X1), 7.974 0.825119 0.813714 61.9002 b0 = 445.0583, p<0.001

Income (X2), (p< 0.001) b1 = 31.79566, p<0.001 and b2 = 0.003698, p<0.001

Kids (X3) b3 = 45.69131, p<0.001

Se = standard error of the estimate

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

1. Write and explain the regression equation.
2. Explain the performance metrics of the regression equation using the table above.
3. 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= b0 + b1x1+ b2x2+ b3x3

=>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.

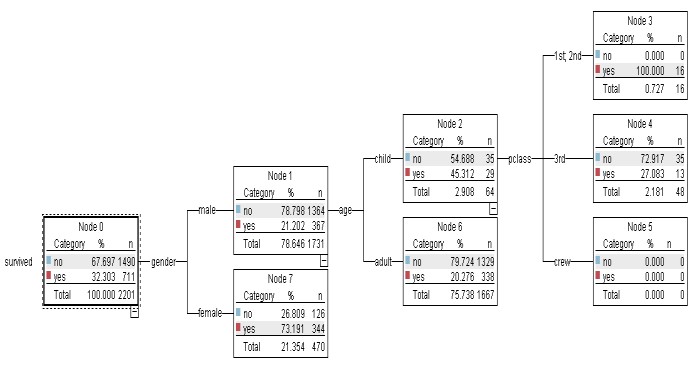
c) the regression equation is donations = 445.0583 + 31.79566\*Education + 0.003698\*Income + 45.69131\*Kids = 445.0583 + 31.79566\*16 + 0.003698\*90 + 45.69131\*2 = 1045.5043 units

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:



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.

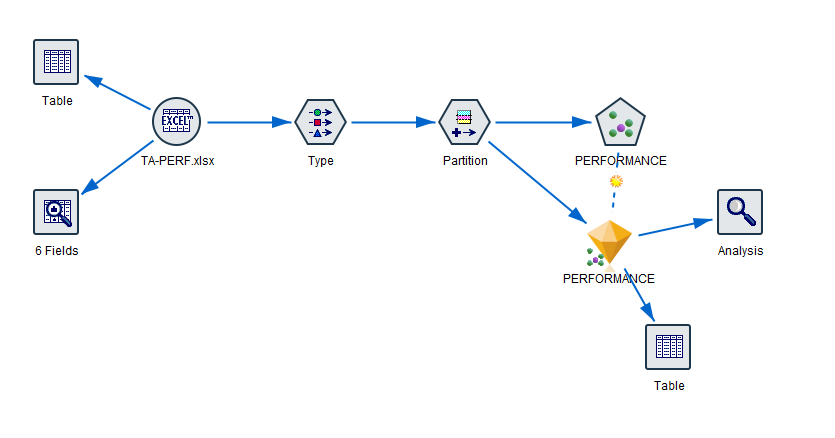
Attribute Information:

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.

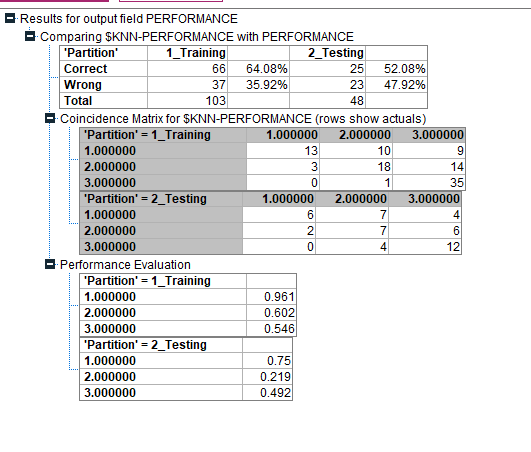
(Use SPSS Modeler to complete this question)

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



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.

