CSC 400, Exam #1, At Home, Spring 2024 Name KEY

Instructions: This exam is in two parts: Part I is to be completed partly at home using the materials posted in the course for the at-home portion and you will answer questions about that work during the in-class portion of the exam; Part II is to be completed entirely in class. You may not use cell phones, and you may only access internet resources you are specifically directed to use.

At home, prepare for questions in Part I using R. Complete the calculations noted below. You will be asked for additional analysis and interpretation of this data in the in-class portion of the test. Print out the results of your analysis and code, and bring the pages with you to the exam. You will submit all this work along with the in-class exam.

Use the data on wine to complete the following tasks after importing the data **400exam1data.xlsx** in the file into R.

1. Create a correlation table of the variables for both Pearson and Spearman correlations. Make a correlation plot and a pairplot of the data set.

2. Create a classification model for the data set (you do not need to separate the data into test and training sets for this) using the following algorithms: Decision tree, KNN and ~~naive Bayes~~ change LDA

3. For each model identified above, create a confusion matrix, and create appropriate model or diagnostic graphs.

Import the dataset **PimaIndiansDiabetes** from the mlbench package.

4. Create an SVM model of the classification task. (You do not need to separate into testing and training sets for this task.)

5. Create a simple neural network model of the classification task.

6. Create a model for the classification task using XGBoost.

7. Create a confusion matrix for each of the models above. Create appropriate diagnostic graphs.

Import the dataset **IncomeESL** from the arules package.

8. Perform association rule mining using the Apriori algorithm on the dataset. Include in your analysis the top 5 rules identified by lift, and by support. You'll be asked about these during the in-class portion of the exam. Be sure to include in your analysis appropriate visualizations.
remove ethnic classification

Neural networks:

9. For this exam, be prepared to discuss:
   a. General applications of neural networks
   b. Convolutional Neural Networks, Long Short-term Memory Networks, Graph Neural Networks

CSC 400, Exam #1, Spring 2024 Name KEY

Instructions: Answer each question thoroughly. For questions in Part 1, use the work you did at home to answer the questions. Be sure to answer each part of each question. In Part 2, report exact answers unless directed to round.

Part I:
Use the work you did at home to answer these questions the wine dataset.

1. Based on your correlation table, which two variables have the highest Pearson correlation? What is the correlation value?

   Flavanoids vs. Total_Phenols

   0.86456

2. Based on your correlation table (or graphs), which two variables have the highest Spearman correlation? What is the correlation value?

   the same

   0.87940437

3. Describe how your correlation plots for the Pearson and Spearman correlations differ, qualitatively. What do you notice about how the plots differ?

   answers may vary
   The sequence of variables is plotted differently

4. In your pairplot, do any of the scatterplots appear to be particularly nonlinear? If so, which pair? Describe the shape and give the correlation value. If you were building a model of the data, is there a transformation you could make to try to improve the fit based on the scatterplot?

   answers may vary

   Flavanoids vs. Color intensity looks possibly quadratic.
   Could try transforming the variable w/ square or square root
   to test for a better match

5. Does the pairplot suggest that there may be outliers in the data that may need to be analyzed and removed? If so, in which variable(s)?

   Yes, several variables show outliers such as Ash
   (also Magnesium, Hue and Flavanoids)

6. For your decision tree, what rule makes the first split?

Prolnie ≥ 755

7. For your decision tree model, what is the accuracy of the model?

93.8%

167/178

8. For your KNN model, which value of k produced the highest accuracy?

unscaled highest w/ K=6

scaled highest w/ k=5 (not only value that gives this accuracy)

9. What was the confusion matrix for k=5 (for KNN)?

noscaling

$$\begin{bmatrix} 51 & 3 & 5 \\ 5 & 53 & 13 \\ 1 & 11 & 36 \end{bmatrix}$$

w/ scaling

$$\begin{bmatrix} 59 & 0 & 0 \\ 2 & 67 & 2 \\ 0 & 0 & 48 \end{bmatrix}$$

10. Did rescaling the variable improve the accuracy of your model? Why or why not?

yes for k=5. the accuracy improved from 78.65% to 97.75%

11. After applying the LDA algorithm to the data, what was the accuracy of that classification model?

100%

12. Given the available data, comparing the Decision Tree model, the KNN model and the LDA model, which model would you choose and why? You may want to consider more than accuracy, for example, interpretation may be important if the goal is to understand the classification scheme.

answers may vary

in this context, the LDA model produced the most accuracy on the original data. Ideally, it should be tested against another data set (test data) to be sure.

If interpretation of the model matters, the decision tree did a pretty good jobs for less computations.

For the questions that follow, use the your analysis of the Pima Indians Diabetes data.

13. In your SVM model with a linear kernel (the default), what was the accuracy achieved with this model?

77.34%

14. Were you able to improve the accuracy with a nonlinear kernel? If so, which one, and what was the improvement?

a polynomial kernel w/ defaults improved to 80.46%
radial basis produced an error, Signoid did worse.
may be able to achieve better % w/ parameters adjustments

15. Compare the results of the SVM model with your simple neural network model. Was the accuracy better, worse or similar?

it was similar, 78.25% vs. 77.34% for linear SVM
a little worse than polynomial kernel

16. What is the confusion matrix you obtained from the XGBoost model?

$$\begin{bmatrix} 486 & 14 \\ 45 & 223 \end{bmatrix}$$

17. Of the three models applied to the diabetes data, which model performed the best and why?

XGBoost did the best, by far w/ 92.32% accuracy

The questions that follow refer to the Income data and association rule analysis.

18. Based on support, what is the top rule? What are the lift and support values?

{marital status = single} ⇒ {dual incomes = not married}

support 0.406
lift 1.65

Part II:

19. Give three examples of applications for neural networks aside from regression and classification.

answers may vary

recommendation systems
NLP
deep learning for unstructured data

20. Convolutional neural networks are particularly useful for which types of applications? Give at least two examples.

image classification
object detection

21. Long Short-Term Memory networks are a type of Recurrent Neural Network (RNN) designed to address what type of issue? To what kinds of problems are they typically applied?

addresses vanishing gradient problem
applied to NLP and time series forecasts

22. What kind of data or structures are graph neural networks designed to work best on? Give examples.

Networks: social networks,
citation networks,
recommendation systems

23. What are three factors to consider when choosing a neural network design?

answers may vary

nature of the problem
data complexity
computational resources

24. Some classification methods are designed around binary classification, such as logistic regression. If you wanted to apply such a model to data with three or more classes, such as the wine dataset from the at-home portion of the exam, what would you need to do? Describe the process (in some detail).

you would to split the data into two classes
say A vs. B&C, Then for anything classified
in the B/C group apply a binary classification
again.

25. What advantages and disadvantages are there to applying ensemble methods to a classification (or regression) problem? Give at least two of each.

advantages
improve accuracy
reduce over fitting

disadvantages
more computationally expensive
more difficult to interpret or explain

26. How do boosting and bagging differ from each other? Explain both.

bagging tries to reduce the variance of a model
boosting tries to reduce bias

answers may vary

27. What simple assumption does Naïve Bayes make that earns it the name "naïve"?

assumes independence of variables

28. K-Means is technically a clustering algorithm that is often used for classification. Explain how that works and why it is different from traditional classification techniques.

clustering is unsupervised learning rather than supervised. the class labels are not used in the model. But we can select the clusters to equal the # of classes and then manually identify which cluster is equivalent to which class and build a confusion matrix from there.

29. What is the main difference between data mining and machine learning? What do they have in common?

The main difference is the purpose
most techniques are shared

30. What is the purpose of the Gini index in a decision tree model?

it is used to determine best splits at nodes

31. In association rule mining, what is the purpose of a parallel coordinates plot?

allows for a visual way of comparing rules across various demensions

CSC 400 Exam #1 At-home analysis

Pearson

	Alcohol	Malic_Acid	Ash	Ash_Alcalinity	Magnesium	Total_Phenols
Alcohol	1.00000000	0.09439694	0.211544596	-0.31023514	0.27079823	0.28910112
Malic_Acid	0.09439694	1.00000000	0.164045470	0.28850040	-0.05457510	-0.33516700
Ash	0.21154460	0.16404547	1.000000000	0.44336719	0.28658669	0.12897954
Ash_Alcalinity	-0.31023514	0.28850040	0.443367187	1.00000000	-0.08333309	-0.32111332
Magnesium	0.27079823	-0.05457510	0.286586691	-0.08333309	1.00000000	0.21440123
Total_Phenols	0.28910112	-0.33516700	0.128979538	-0.32111332	0.21440123	1.00000000
Flavanoids	0.23681493	-0.41100659	0.115077279	-0.35136986	0.19578377	0.86456350
Nonflavanoid_Phenols	-0.15592947	0.29297713	0.186230446	0.36192172	-0.25629405	-0.44993530
Proanthocyanins	0.13669791	-0.22074619	0.009651935	-0.19732684	0.23644061	0.61241308
Color_Intensity	0.54636420	0.24898534	0.258887259	0.01873198	0.19995001	-0.05513642
Hue	-0.07174720	-0.56129569	-0.074666889	-0.27395522	0.05539820	0.43368134
OD280/OD315	0.07234319	-0.36871043	0.003911231	-0.27676855	0.06600394	0.69994936
Proline	0.64372004	-0.19201056	0.223626264	-0.44059693	0.39335085	0.49811488

	Flavanoids	Nonflavanoid_Phenols	Proanthocyanins	Color_Intensity	Hue
Alcohol	0.2368149	-0.1559295	0.136697912	0.54636420	-0.07174720
Malic_Acid	-0.4110066	0.2929771	-0.220746187	0.24898534	-0.56129569
Ash	0.1150773	0.1862304	0.009651935	0.25888726	-0.07466689
Ash_Alcalinity	-0.3513699	0.3619217	-0.197326836	0.01873198	-0.27395522
Magnesium	0.1957838	-0.2562940	0.236440610	0.19995001	0.05539820
Total_Phenols	0.8645635	-0.4499353	0.612413084	-0.05513642	0.43368134
Flavanoids	1.0000000	-0.5378996	0.652691769	-0.17237940	0.54347857
Nonflavanoid_Phenols	-0.5378996	1.0000000	-0.365845099	0.13905701	-0.26263963
Proanthocyanins	0.6526918	-0.3658451	1.000000000	-0.02524993	0.29554425
Color_Intensity	-0.1723794	0.1390570	-0.025249931	1.00000000	-0.52181319
Hue	0.5434786	-0.2626396	0.295544253	-0.52181319	1.00000000
OD280/OD315	0.7871939	-0.5032696	0.519067096	-0.42881494	0.56546829
Proline	0.4941931	-0.3113852	0.330416700	0.31610011	0.23618345

	OD280/OD315	Proline
Alcohol	0.072343187	0.6437200
Malic_Acid	-0.368710428	-0.1920106
Ash	0.003911231	0.2236263
Ash_Alcalinity	-0.276768549	-0.4405969
Magnesium	0.066003936	0.3933508
Total_Phenols	0.699949365	0.4981149
Flavanoids	0.787193902	0.4941931
Nonflavanoid_Phenols	-0.503269596	-0.3113852
Proanthocyanins	0.519067096	0.3304167
Color_Intensity	-0.428814942	0.3161001
Hue	0.565468293	0.2361834
OD280/OD315	1.000000000	0.3127611
Proline	0.312761075	1.0000000

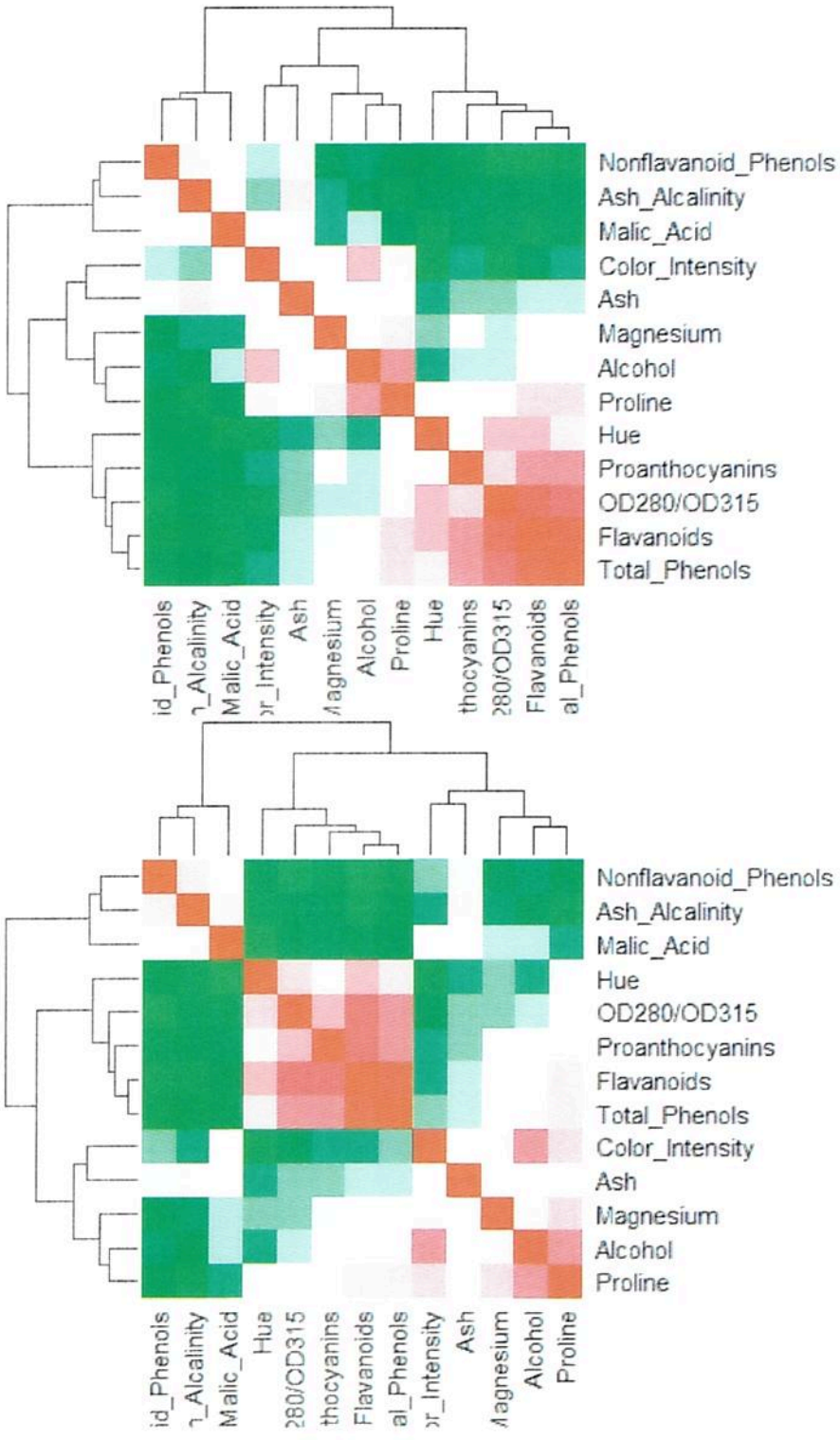
Spearman

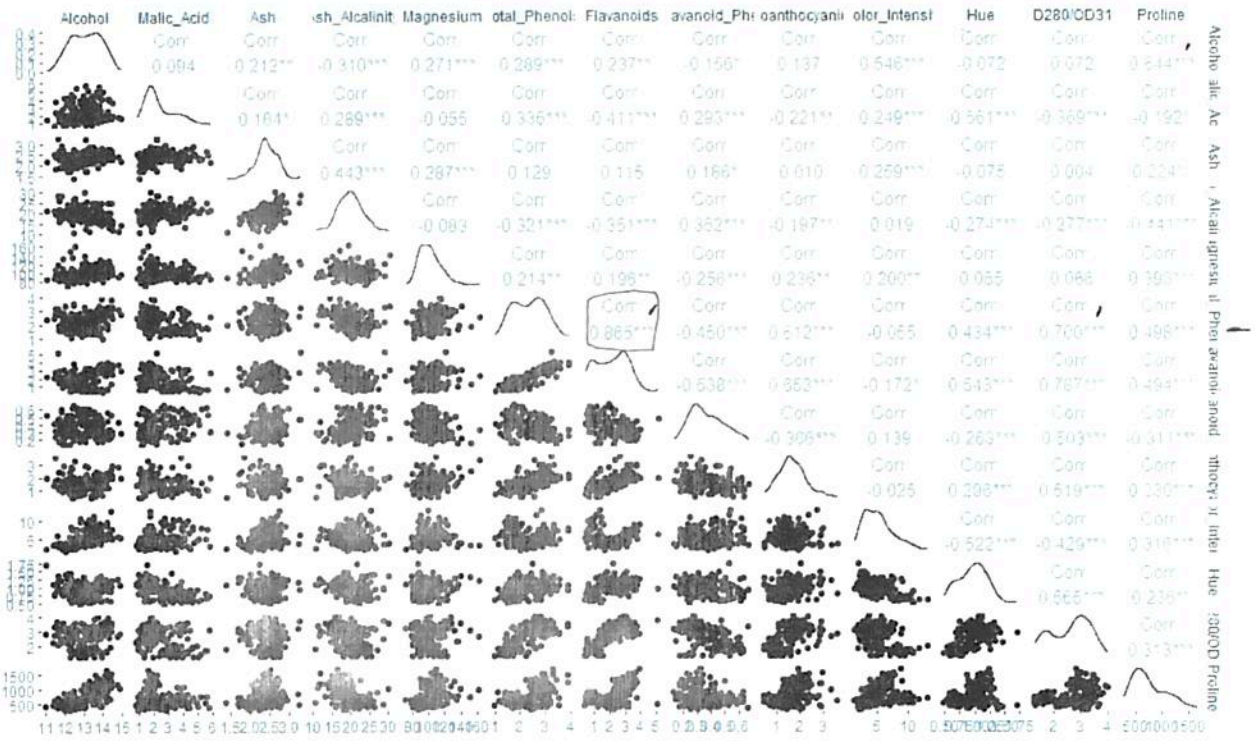
	Alcohol	Malic_Acid	Ash	Ash_Alcalinity	Magnesium	Total_Phenols
Alcohol	1.00000000	0.14043018	0.24372221	-0.30659789	0.36550338	0.31092007
Malic_Acid	0.14043018	1.00000000	0.23067358	0.30406913	0.08018776	-0.28022473
Ash	0.24372221	0.23067358	1.00000000	0.36637364	0.36148788	0.13219338

Ash_Alcalinity	-0.30659789	0.30406913	0.36637364	1.00000000	-0.16955822	-0.37665712
Magnesium	0.36550338	0.08018776	0.36148788	-0.16955822	1.00000000	0.24641696
Total_Phenols	0.31092007	-0.28022473	0.13219338	-0.37665712	0.24641696	1.00000000
Flavanoids	0.29473988	-0.32520214	0.07879581	-0.44376999	0.23316660	0.87940437
Nonflavanoid_Phenols	-0.16220710	0.25523566	0.14558274	0.38938987	-0.23678609	-0.44801307
Proanthocyanins	0.19273419	-0.24482493	0.02438423	-0.25369530	0.17364700	0.66668937
Color_Intensity	0.63542453	0.29030671	0.28304684	-0.07377598	0.35702874	0.01116179
Hue	-0.02420284	-0.56026489	-0.05018326	-0.35250702	0.03609452	0.43945748
OD280/OD315	0.10305021	-0.25518507	-0.00749991	-0.32588976	0.05696275	0.68720700
Proline	0.63357986	-0.05746615	0.25316348	-0.45608973	0.50757486	0.41946986

	Flavanoids	Nonflavanoid_Phenols	Proanthocyanins	Color_Intensity	Hue
Alcohol	0.29473988	-0.16220710	0.19273419	0.63542453	-0.02420284
Malic_Acid	-0.32520214	0.25523566	-0.24482493	0.29030671	-0.56026489
Ash	0.07879581	0.14558274	0.02438423	0.28304684	-0.05018326
Ash_Alcalinity	-0.44376999	0.38938987	-0.25369530	-0.07377598	-0.35250702
Magnesium	0.23316660	-0.23678609	0.17364700	0.35702874	0.03609452
Total_Phenols	0.87940437	-0.44801307	0.66668937	0.01116179	0.43945748
Flavanoids	1.00000000	-0.54389740	0.73032169	-0.04291039	0.53543014
Nonflavanoid_Phenols	-0.54389740	1.00000000	-0.38462885	0.05963853	-0.26781255
Proanthocyanins	0.73032169	-0.38462885	1.00000000	-0.03094715	0.34279465
Color_Intensity	-0.04291039	0.05963853	-0.03094715	1.00000000	-0.41852176
Hue	0.53543014	-0.26781255	0.34279465	-0.41852176	1.00000000
OD280/OD315	0.74153289	-0.49494998	0.55403138	-0.31751560	0.48545434
Proline	0.42990441	-0.27011194	0.30824932	0.45709642	0.20774049

	OD280/OD315	Proline
Alcohol	0.10305021	0.63357986
Malic_Acid	-0.25518507	-0.05746615
Ash	-0.00749991	0.25316348
Ash_Alcalinity	-0.32588976	-0.45608973
Magnesium	0.05696275	0.50757486
Total_Phenols	0.68720700	0.41946986
Flavanoids	0.74153289	0.42990441
Nonflavanoid_Phenols	-0.49494998	-0.27011194
Proanthocyanins	0.55403138	0.30824932
Color_Intensity	-0.31751560	0.45709642
Hue	0.48545434	0.20774049
OD280/OD315	1.00000000	0.25326568
Proline	0.25326568	1.00000000

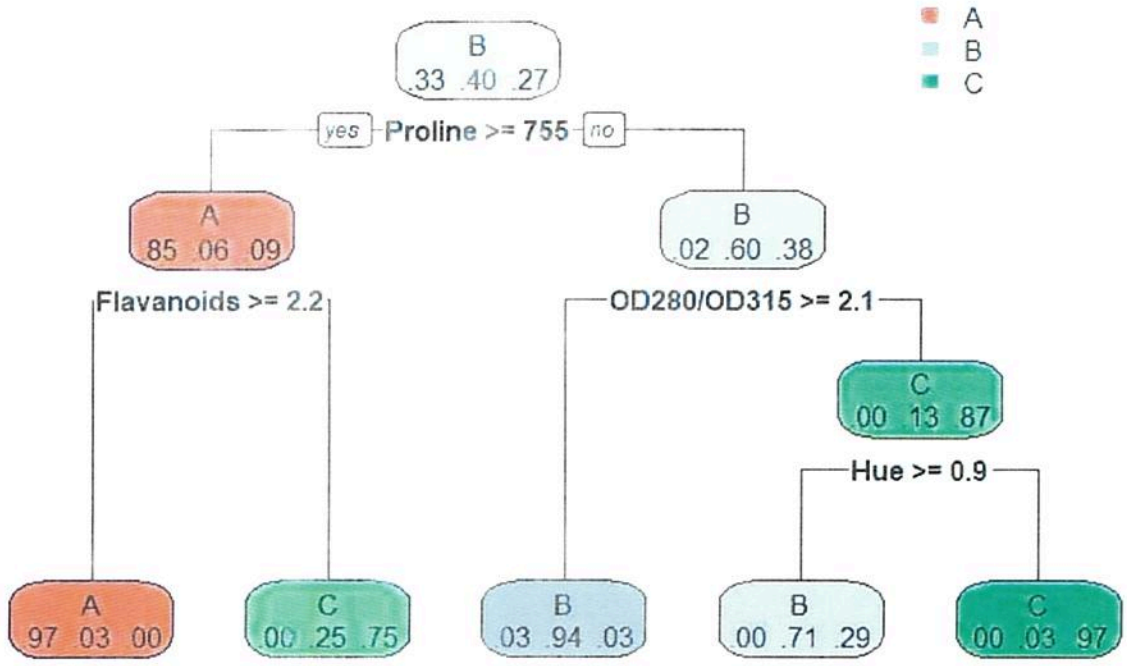

Nonflavanoid_Phenols
Ash_Alcalinity
Malic_Acid
Color_Intensity
Ash
Magnesium
Alcohol
Proline
Hue
Proanthocyanins
OD280/OD315
Flavanoids
Total_Phenols
id_Phenols
Alcalinity
Malic_Acid
r_Intensity
Ash
Magnesium
Alcohol
Proline
Hue
thocyanins
280/OD315
Flavanoids
al_Phenols
Nonflavanoid_Phenols
Ash_Alcalinity
Malic_Acid
Hue
OD280/OD315
Proanthocyanins
Flavanoids
Total_Phenols
Color_Intensity
Ash
Magnesium
Alcohol
Proline
id_Phenols
Alcalinity
Malic_Acid
Hue
280/OD315
thocyanins
Flavanoids
al_Phenols
r_Intensity
Ash
Magnesium
Alcohol
Proline



Type

A B C
0.3314607 0.3988764 0.2696629

Decision Tree



```
    A  B  C
A 57  2  0
B  2 66  3
C  0  4 44
```

167/178= Accuracy (0.938…)

KNN
No scaling, k=5

```
   test1
    A  B  C
A 51  3  5
B  5 53 13
C  1 11 36
```

140/178 = Accuracy (0.7865…)

Highest with k=6

```
   test1
    A  B  C
A 55  1  3
B  4 55 12
C  2 10 36
```

Accuracy: 82.02%

With scaling, k=5

```
   test2
    A  B  C
A 59  0  0
B  2 67  2
C  0  0 48
```

174/178=Accuracy (0.9775…)

K=5 appears to be the highest (although there is at least one other value of k that gives the same accuracy on this data).

LDA

```
             Predicted Group
Actual Group  A  B  C
           A 59  0  0
           B  0 71  0
           C  0  0 48
```

Accuracy: 100%

Pima Indians Data
Linear:

```
   y_pred
     0   1
0 442  58
1 116 152
```

594/768 = Accuracy (0.7734…)

Polynomial kernel (with defaults):

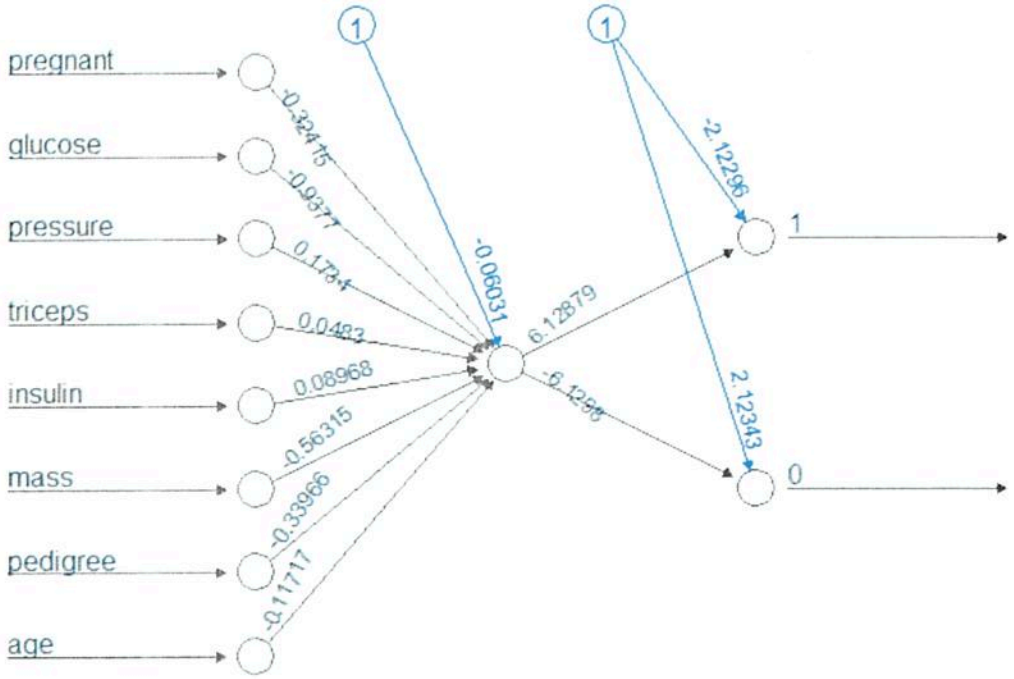
```
   y_pred
```

```
     0   1
  0 485  15
  1 135 133
```

618/768 = Accuracy (0.8046…)

Radial basis generates an error, sigmoid performs worse than linear; it may b
e possible to improve performance of the polynomial kernel by adjusting the d
efault parameters

Neural network



Error: 717.211965 Steps: 714

```
   prediction
      0   1
  0 435  65
  1 102 166
```

601/768 = accuracy (0.7825…)

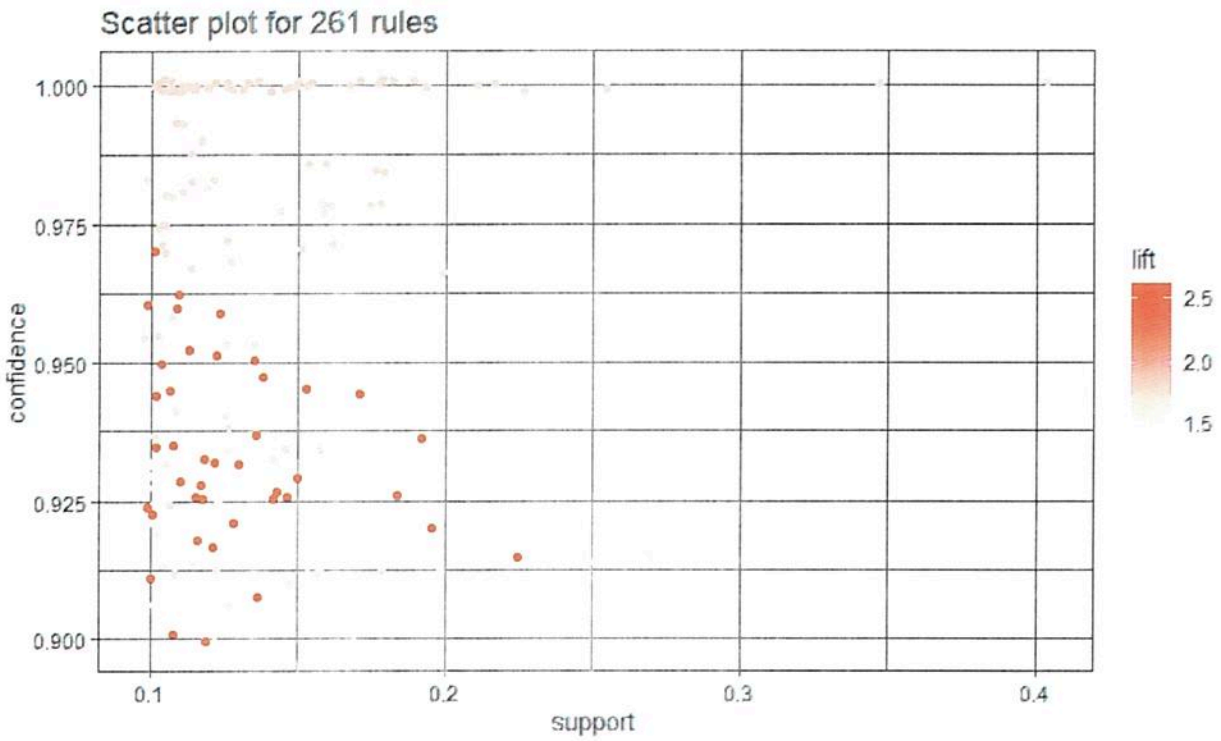
XGBoost

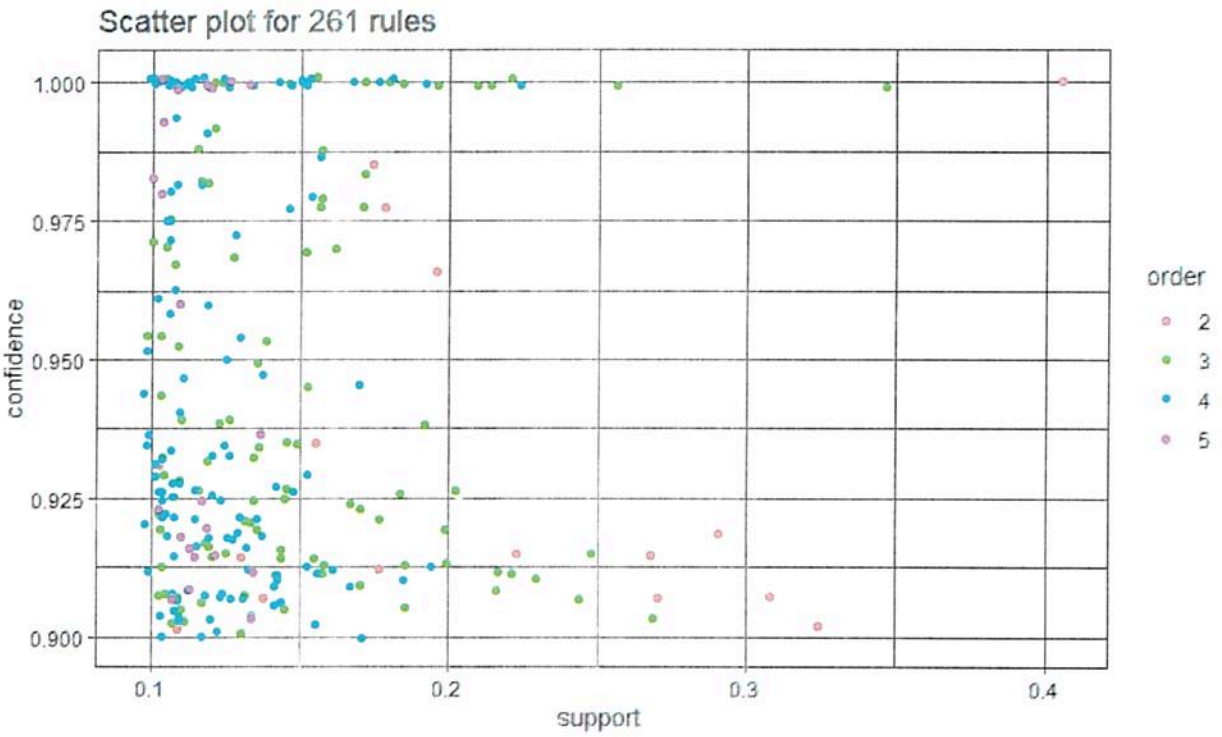
```
                Length Class              Mode
handle               1 xgb.Booster.handle externalptr
raw              61311 -none-             raw
niter                1 -none-             numeric
evaluation_log       2 data.table         list
call                14 -none-             call
params               2 -none-             list
callbacks            2 -none-             list
feature_names        8 -none-             character
nfeatures            1 -none-             numeric
```

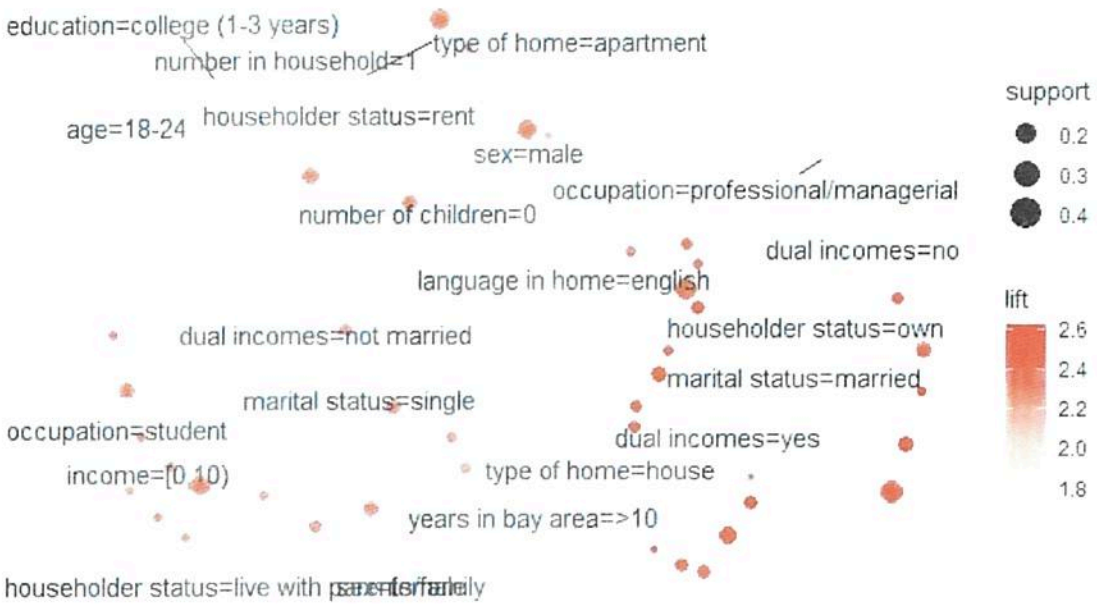
Confusion Matrix and Statistics

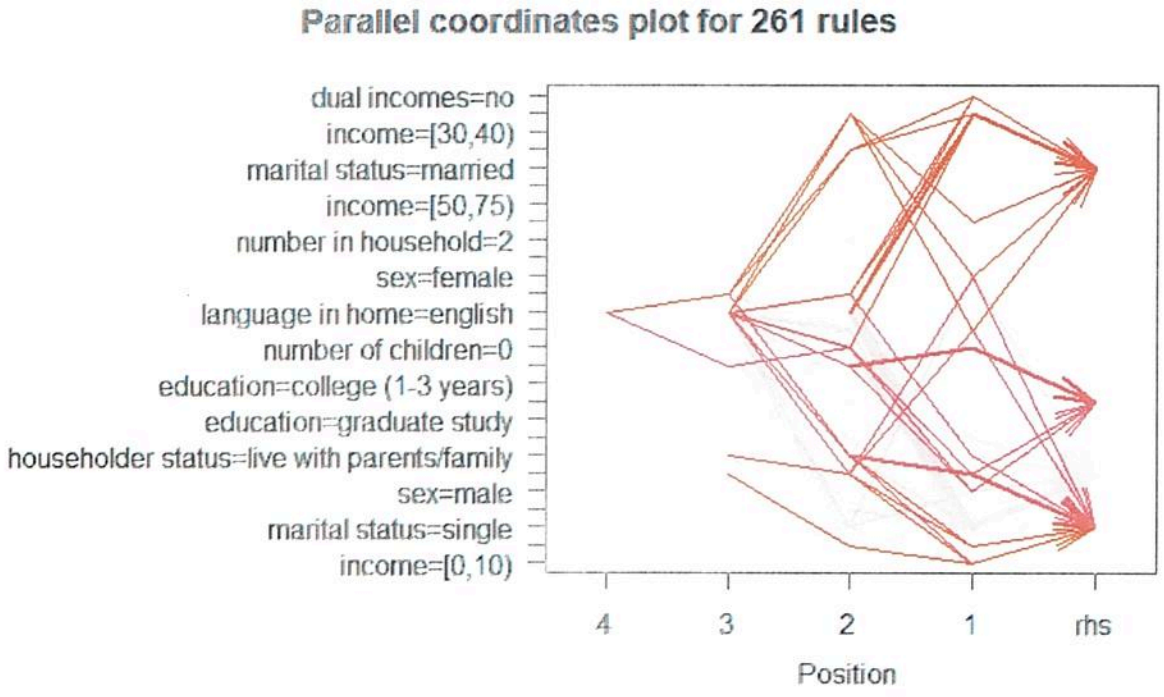
```
          Reference
```

```
Prediction   0   1
         0 486  14
         1  45 223

               Accuracy : 0.9232
                 95% CI : (0.902, 0.941)
    No Information Rate : 0.6914
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.8263

 Mcnemar's Test P-Value : 9.397e-05

            Sensitivity : 0.9153
            Specificity : 0.9409
         Pos Pred Value : 0.9720
         Neg Pred Value : 0.8321
             Prevalence : 0.6914
         Detection Rate : 0.6328
   Detection Prevalence : 0.6510
      Balanced Accuracy : 0.9281

       'Positive' Class : 0


Association rules/Apriori
By lift:

# A tibble: 5 × 6
  rules                                                                 support confidenc
e coverage  lift count
  <chr>                                                                   <dbl>      <dbl
>    <dbl> <dbl> <int>
  {dual incomes=no,householder status=own} => {marital sta…   0.102      0.97
1    0.105  2.62   914
  {years in bay area=>10,dual incomes=yes,type of home=hou…   0.100      0.96
1    0.104  2.59   902
  {dual incomes=yes,householder status=own,type of home=ho…   0.110      0.96
0    0.114  2.59   988
  {dual incomes=yes,householder status=own,type of home=ho…   0.121      0.95
9    0.126  2.59  1088
  {dual incomes=yes,householder status=own,language in hom…   0.125      0.95
1    0.131  2.56  1122

By support:

# A tibble: 5 × 6
  rules                                                                 support confidenc
e coverage  lift count
  <chr>                                                                   <dbl>      <dbl
>    <dbl> <dbl> <int>
  {marital status=single} => {dual incomes=not married}       0.406      1
0.406  1.65  3654
  {marital status=single,language in home=english} => {dua…   0.347      1
0.347  1.65  3120
  {householder status=own} => {language in home=english}      0.327      0.90
2    0.362  1.04  2937
  {education=college (1-3 years)} => {language in home=eng…   0.309      0.90
7    0.341  1.05  2782
  {occupation=professional/managerial} => {language in hom…   0.288      0.92
0    0.314  1.06  2593
```


Scatter plot for 261 rules
1.000
0.975
0.950
0.925
0.900
confidence
0.1
0.2
0.3
0.4
support
lift
2.5
2.0
1.5


Scatter plot for 261 rules
1.000
0.975
0.950
0.925
0.900
confidence
0.1
0.2
0.3
0.4
support
order
2
3
4
5


education=college (1-3 years)
type of home=apartment
number in household=1
age=18-24
householder status=rent
sex=male
occupation=professional/managerial
number of children=0
dual incomes=no
language in home=english
householder status=own
dual incomes=not married
marital status=married
marital status=single
occupation=student
dual incomes=yes
income=[0,10)
type of home=house
years in bay area=>10
householder status=live with parents/family
support
0.2
0.3
0.4
lift
2.6
2.4
2.2
2.0
1.8


Parallel coordinates plot for 261 rules
dual incomes=no
income=[30,40)
marital status=married
income=[50,75)
number in household=2
sex=female
language in home=english
number of children=0
education=college (1-3 years)
education=graduate study
householder status=live with parents/family
sex=male
marital status=single
income=[0,10)
4
3
2
1
rhs
Position