There are ten questions. Each question may cover several topics. The exam is Due on February 16th. Submit it online. Each question 1 to 10 is worth 10 points. It should take you from 2 to 4 hours.

In questions 3, 4, 6, 7A, include all r code or other methods used. For example, it is possible to use Excel to carry out some of the analysis. If you use Excel or something other than R take a screen shot of the setup and the results. Basically, show your work.

1. **A hospital in interested in learning about the risk of readmission for people who were treated for a specific chronic ailment. They want to know what contributes to an individual’s risk of readmission. Risk of readmission is based on the likelihood that the person will come back to the hospital within a short period of time. There are several data mining tasks that are relevant in this situation. Discuss the relevant ones and how the analysis may be of value to the hospital.**

* **Develop an Understanding of the purpose of Data Mining Project**
* **Obtain the Dataset to be used in analysis:** After gathering the full data initially, the team can filter the patients’ data with those who have persistent ailments.
* **Explore, Clean and Preprocess the Data:** Exploratory Analysis will help the team to locate extremely necessary cases. This process helps determine whether a certain patient’s data can be eliminated or not.
* **Reduce the data dimension, if necessary:** Splitting the data into a basic 60:40 split can help ease the process as datasets can be quite large and partitioning may help achieve better results.
* **Determine the Data Mining Task:** The team can do this by further creating specific/necessary questions or topics of interest for creating this model.
* **Partition the Data (for supervised tasks):** Maybe categorizing the datasets by department or by illness of patients the datasets can be trained, validated and tested.
* **Choose the data mining techniques to be used:** Whether it is regression, clustering or other methods, the team should select the technique if the dependent variables are highly correlated. Different Regression methods can have a base by analyzing the correlation between variables.
* **Use Algorithms to perform the task:** Validation of the dataset can be quite useful in this step. Since it can be a very iterative process, Validation would provide feedback on the performance of the algorithms using different variations.
* **Interpret the results of the algorithms**
* **Deploy the Model**

1. **A. Describe a situation where you want to keep outliers. Describe one where you don’t. Give an example of each.**

**When to Drop an Outlier:** When the data is incorrectly entered and have no means of rectifying it or if it impacts the results of the task at hand, we should drop the outlier. For example: In a dataset of the employees in an office, the age of one employee was recorded as 1. It is impossible to have a one year old in an office. Since the dataset didn’t have names of employees, the employee’s true age could be 21,31,41,51,61. Because of this we can remove this outlier.

**When to keep an Outlier:** We can keep the outlier if we have means to get accurate data for the incorrectly entered data. Using the same example as above, if we knew the name of the employee we could rectify the age and enter the correct age before proceeding with the task at hand.

**B. Missing data can be problematic. There are several ways for handling missing data.**

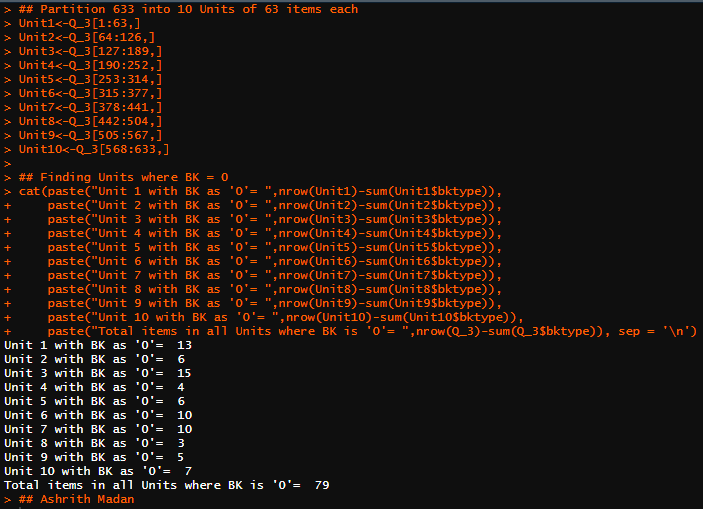
**Given the following table define two *data imputation* methods which we have covered for filling in the missing values.**

**NOTE: You will want to implement your methods and fill in the missing values in the table.**

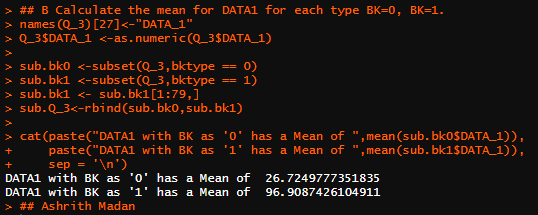
* ***Method 1*:** Imputation Using Mean/Median Values
* ***Method 2*:** Fill the missing values by writing down values of variables closest to it

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **SMBL** | **DATA1** | **DATA2** | **DATA3** | **DATA4** | **DATA5** | **DATA6** | **DATA7** | **DATA8** |
| **AMCVQ** | 42.688 | 1.205 | 2.529 | 51.843 | 83.727 | 293.99 | 317.446 | 225.398 |
| **AMCVQ** | 122.87 | 4.552 | 5.417 | 141.299 | 235.692 | 752.615 | 687.529 | 579.124 |
| **FLMIQ** | 30.316 | 334.278 | 1018.666 | 1494.99 | 1154.597 | 3923.971 | 1598.786 | 949.843 |
| **FLMIQ** | 5.967 | 450.905 | ***1018.666*** | 1587.916 | 1281.084 | 3490.832 | 1554.884 | 820.065 |
| **FLMIQ** | 6.683 | 496.159 | 997.805 | 1728.75 | 1283.719 | 3573.218 | 1539.465 | 838.176 |
| **GRA** | 78.3 | 975.7 | 514.2 | 2228.9 | ***1283.719*** | 6230.6 | 3228.3 | 1730.1 |
| **GRA** | 40.6 | 596.8 | 491.9 | 1681.3 | 2214.2 | 6297.6 | 3154.9 | 1736.1 |
| **GRA** | 68.3 | 831.4 | 376.1 | 1774.9 | 1487.1 | 4945.8 | 3307.9 | 1871.3 |
| **3ABNTQ** | 10.769 | 24.714 | ***376.1*** | 57.87 | 43.906 | 217.751 | 92.565 | 50.138 |
| **3ABNTQ** | 9.101 | 26.922 | 24.752 | 66.764 | ***43.906*** | 201.567 | 113.664 | 53.285 |
| **3ABNTQ** | 14.009 | 22.981 | 21.229 | 66.798 | 138.719 | 184.281 | 105.423 | 42.931 |
| **ENRNQ** | 114.917 | 1576.158 | 111.463 | 2726.907 | 2431.992 | 13238.94 | 11107.18 | 6868.435 |
| **ENRNQ** | 256 | 2255 | 164 | 3979 | 3708 | 16137 | 11348 | 7112 |
| **ENRNQ** | 170 | 2151 | ***164*** | 4669 | 4412 | 23422 | 13742 | 9170 |
| **PTFC** | 49.095 | 68.454 | 327.389 | 460.97 | 318.313 | 1563.586 | 1099.739 | 612.082 |
| **PTFC** | 43.474 | 62.42 | 283.631 | 404.144 | 1553.831 | 1228.061 | 1001.243 | 471.075 |
| **PTFC** | 51.759 | 49.722 | 268.55 | 387.359 | 240.931 | 1020.457 | 309.609 | 287.098 |
| **RVSIQ** | 0.465 | 2.157 | 2.136 | 5.449 | 6.215 | 7.889 | 3.191 | 1.317 |
| **RVSIQ** | 3.063 | 3.412 | ***2.136*** | 10.406 | 5.742 | 14.988 | 3.226 | 1.923 |
| **RVSIQ** | 17.407 | 11.884 | 8.074 | 39.894 | 15.313 | 47.288 | 7.269 | 3.999 |
| **CQB** | 97.863 | 361.198 | 421.806 | 902.867 | 488.422 | 2596.127 | 1892.729 | 1177.823 |
| **CQB** | 96.924 | 296.494 | 428.26 | 846.513 | 612.87 | 2416.79 | 1807.437 | 1071.341 |
| **CQB** | 70.428 | 274.323 | 392.19 | 772.355 | ***612.87*** | 2262.492 | 1806.793 | 1005.606 |

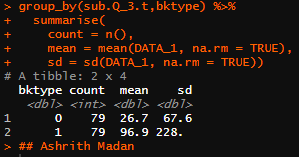
1. **Using the data set “dataforquestion3” included in the Exam module –**
2. **Split the data into 10 equal (roughly) partitions. That means that you will have 633/10 records in each partition. Determine the number of records with a value of BK = 0 in each partition – how many a value of BK = 0 in each partition.**

* ****

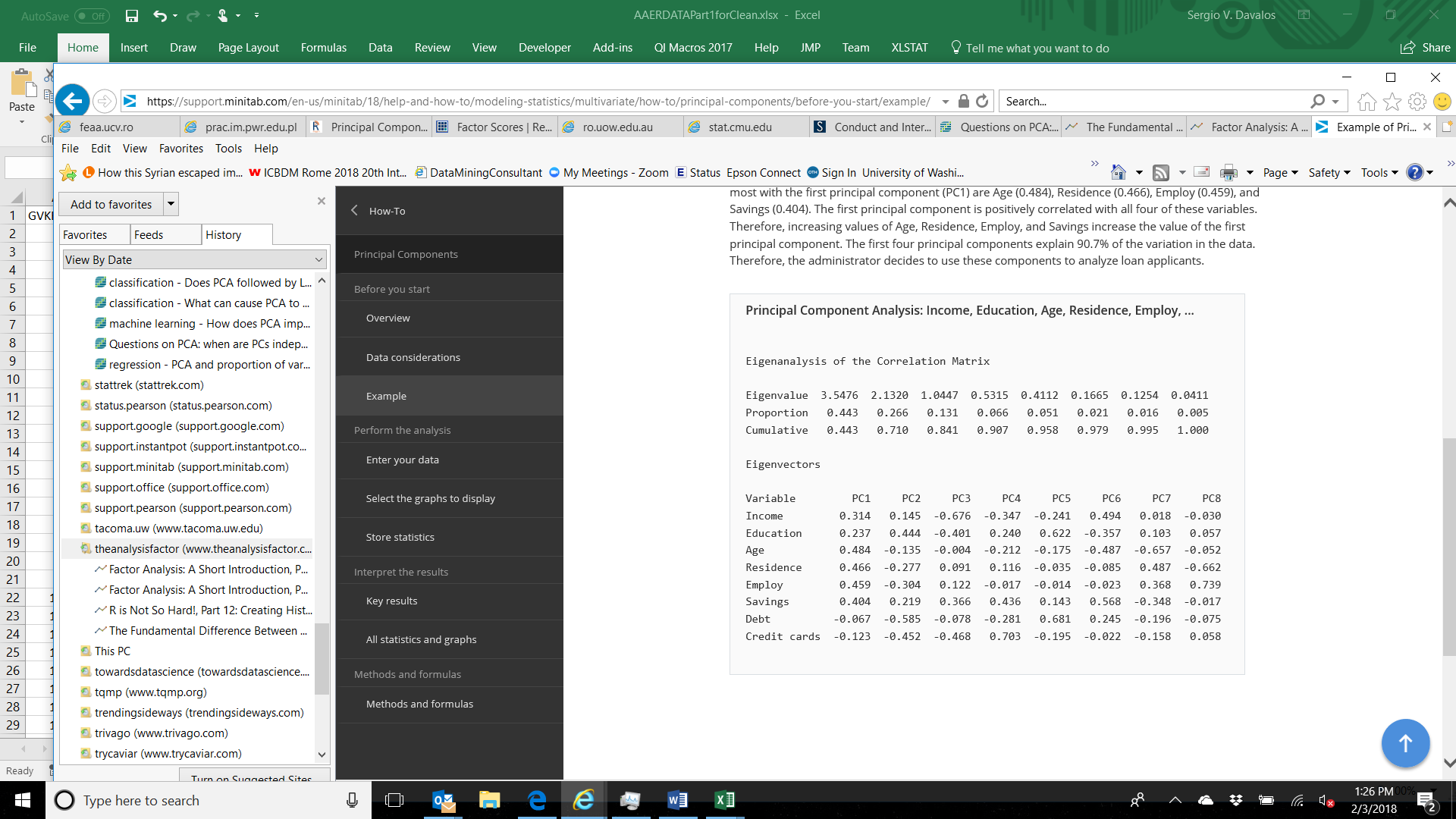
1. **Generate a subset of the data set with all the records that have a value of BK = 0 (should be 79) and an equal number of BK=1 (randomly selected from the remaining records, BK=1). Calculate the mean for DATA1 for each type BK=0, BK=1.**

****

1. **Conduct a t test for difference in the means of the two groups in the subset. Put your results here.**

****

1. **Below is the output from a PCA.**



1. **How many principal components should we use? Explain.**

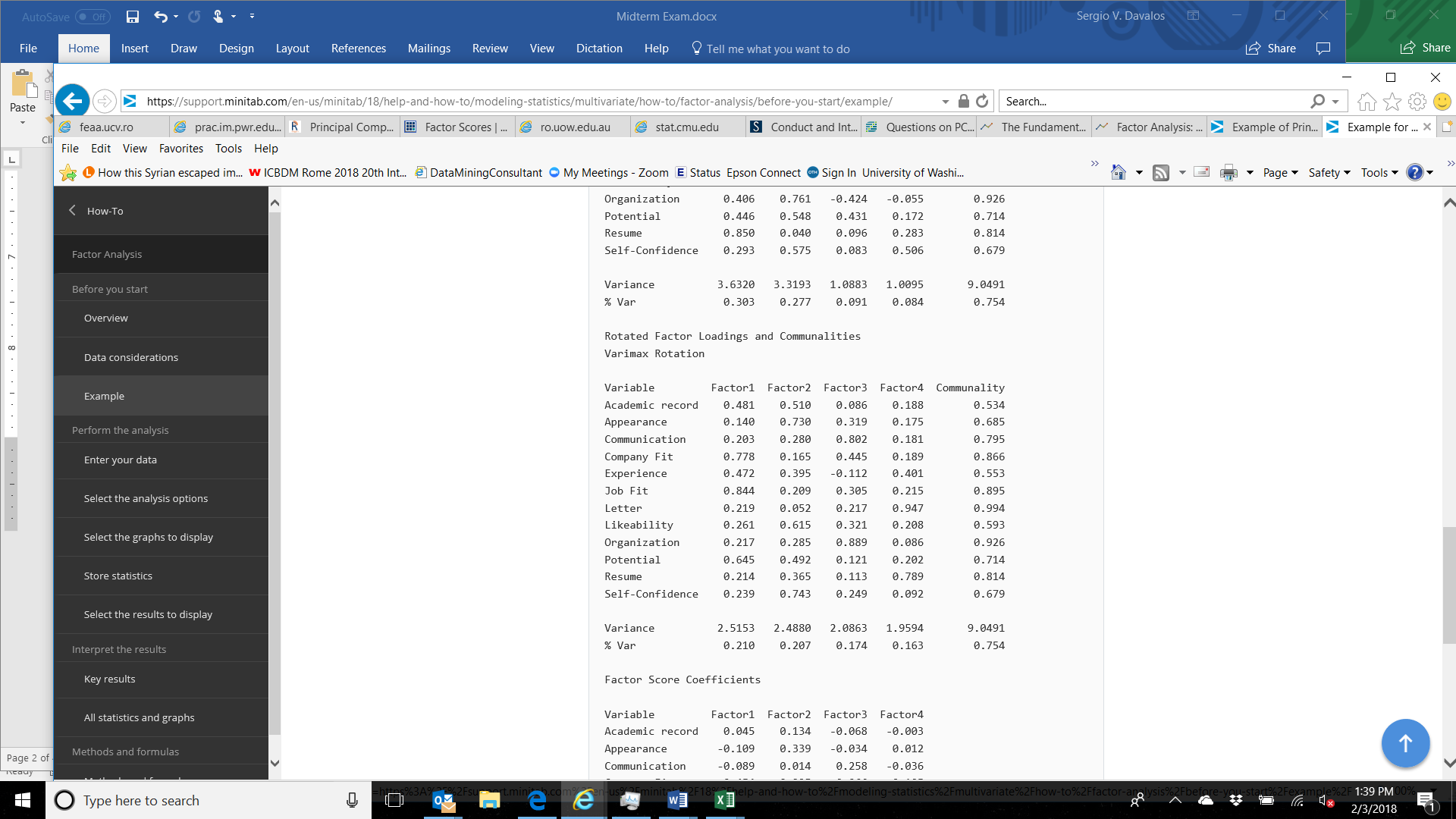
* I would use the first three Principal Components as their Eigenvalues are greater than 1. Having an Eigenvalue of more than 1 shows that they are highly significant. A model without overfitting can be created using these 3 Principal Components. I will consider adding another Principal Component if I find the previous model lacking accuracy.

1. **Discuss the weights for the first four PCs – PC1 to PC4. How do we interpret them?**

* **PC 1 –** This shows a group pf people who show strong positive correlation with Income, Education, Age, Residence, Employ and Savings. It also shows that these people may not be in debt as they are showing negative correlation. This could mean that this group of people are financially sound.
* **PC 2 –** This set of people show negative correlation with Age, Residence, Employ, Debt and Credit Cards. There is slight positive correlation with Income, Education and Savings. This could be adolescents who are still studying and living in rental units.
* **PC 3 –** There is negative correlation with Income, Education, Age, Debt and Credit Cards. But there is a positive correlation with Residence, Employ and Savings. This could also mean that this group of people may be in High School working a part time job and haven’t established line of Credit.
* **PC 4 –** There is negative correlation in Income, Age, Employ and Debt. This could be recently retired as Income and Employ is negative, but they have a strong positive correlation with Savings and Credit Cards.

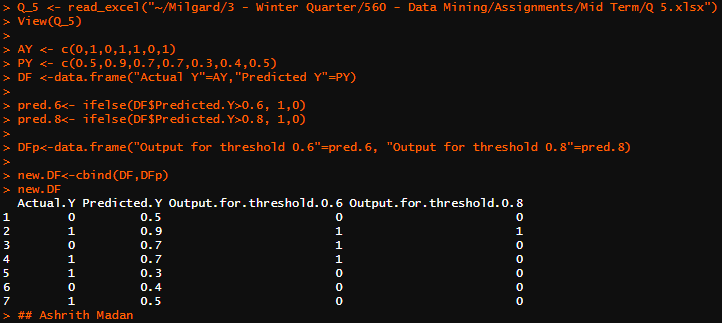
1. **Using the Factor loading table below, identify the latent factors and give them a label. Justify your labels.**

* **Best Fit Factor:** High Scores for Potential, Job it and Company Fit. The Candidate also has decent experience and a good academic record but needs to work on personality.
* **Confidence Factor:** High Scores for Appearance, Likeability and Confidence. Has a good academic record but needs to work on technical Skills.
* **Communication Factor:** Communicates very well and is very organized but needs to improve in a lot of aspects. Has least or no job experience amongst others.
* **Benchmark factor:** Has avery good resume and a recommendation letter**.** Also has some experience but needs to improve on technical skills and personality development.

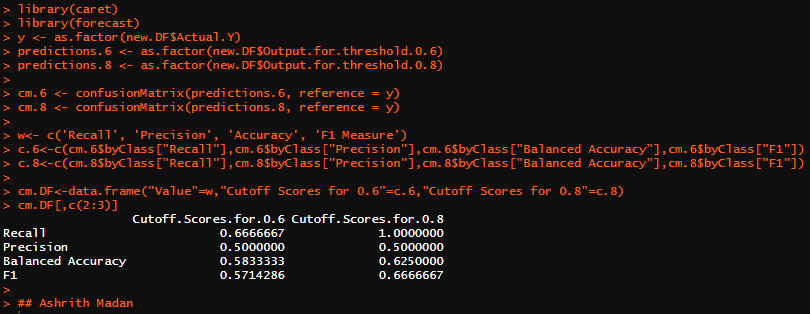


1. **A. Given the table below calculate the confusion matrix for cutoff of 0.6 and 0.8.**

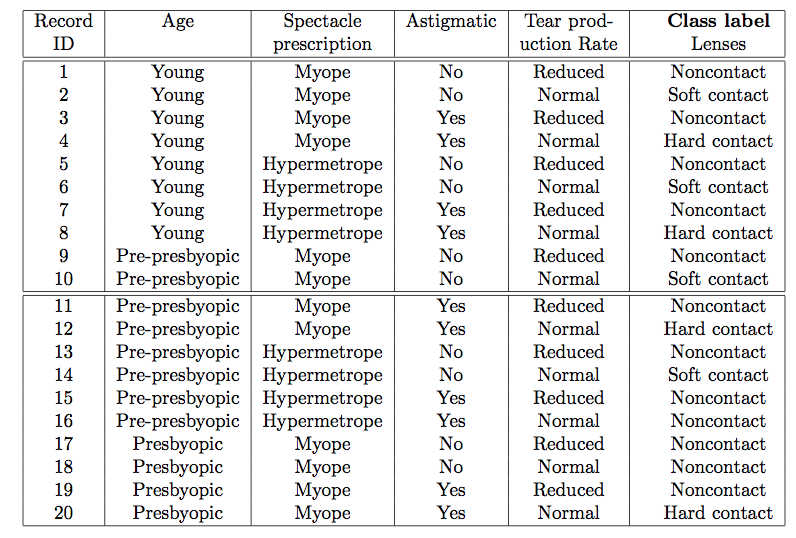
|  |  |  |  |
| --- | --- | --- | --- |
| Actual Y | Predicted y | Output for threshold 0.6 | Output for threshold 0.8 |
| 0 | 0.5 | **0** | **0** |
| 1 | 0.9 | **1** | **1** |
| 0 | 0.7 | **1** | **0** |
| 1 | 0.7 | **1** | **0** |
| 1 | 0.3 | **0** | **0** |
| 0 | 0.4 | **0** | **0** |
| 1 | 0.5 | **0** | **0** |

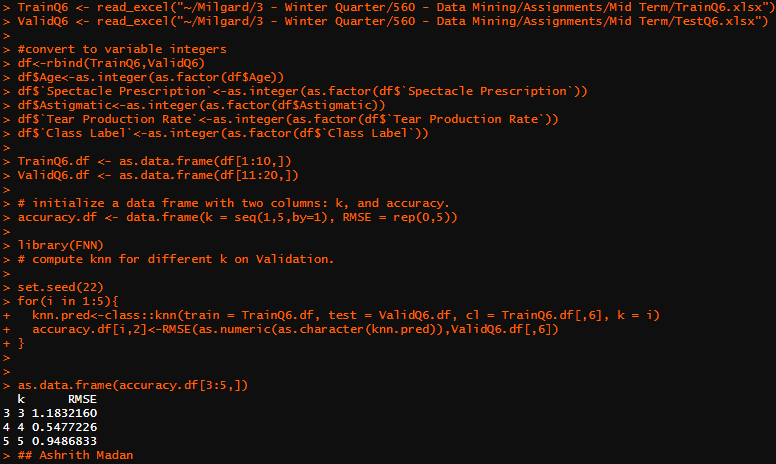


**B. Provide the recall, precision, accuracy, and F1 Measure values for both 0.6 and 0.8.**



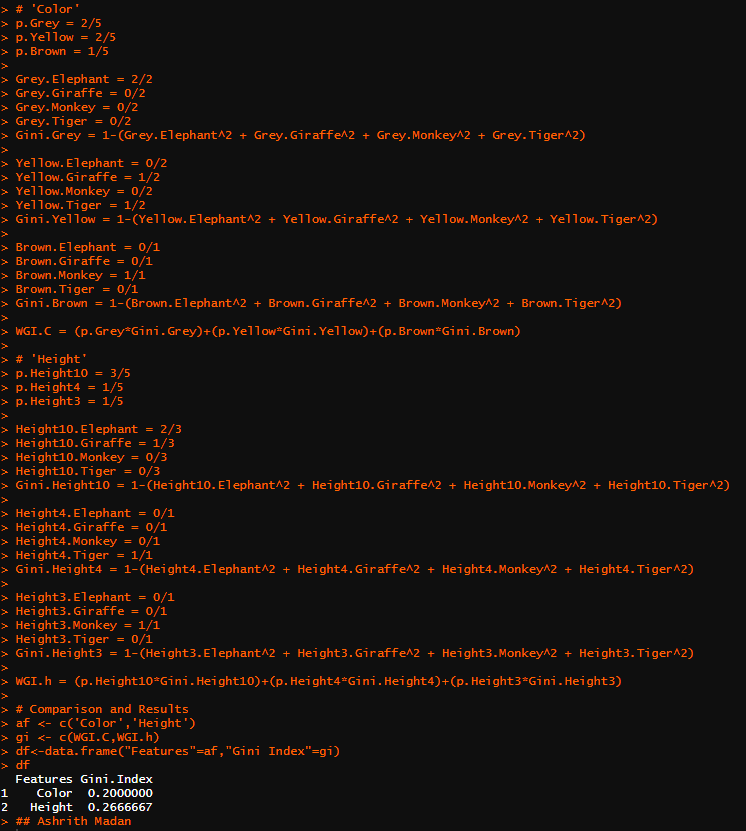
1. **A. The following table below contains 10 records for the training data and 10 for the test. Using KNN identify the test data classes. Determine the K from possible values of 3, 4, and 5 that will give you the best accuracy for the test cases. Note: you will want to convert the variable values to integer to perform KNN.**



* 
* The best accuracy for the test cases can be obtained from the k value of 4 as it would be the best fit among 3,4 & 5.

1. **Given the following table construct the decision tree manually using the Gini impurity measure:**

|  |  |  |
| --- | --- | --- |
| Color | Height | Label |
| Grey | 10 | Elephant |
| Yellow | 10 | Giraffe |
| Brown | 3 | Monkey |
| Grey | 10 | Elephant |
| Yellow | 4 | Tiger |

* 

1. **Say you have 1000 fruits which could be either ‘banana’, ‘orange’ or ‘other’. These are the 3 possible classes of the Y variable.**

**We have data for the following X variables, all of which are binary (1 or 0).**

* **Long**
* **Sweet**
* **Yellow**

**The following table is an example of the data:**

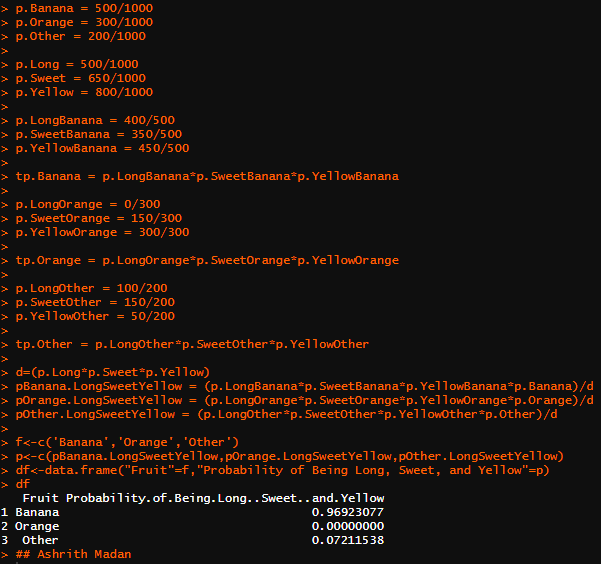
| **Fruit** | **Long (x1)** | **Sweet (x2)** | **Yellow (x3)** |
| --- | --- | --- | --- |
| Orange | 0 | 1 | 0 |
| Banana | 1 | 0 | 1 |
| Banana | 1 | 1 | 1 |
| Other | 1 | 1 | 0 |

**For computing the probabilities, the training data is aggregated to form a counts table like this.**

**[](https://www.machinelearningplus.com/wp-content/uploads/2018/11/05_Naive_bayes_example_new.png)**

1. **If a fruit is long, sweet, and yellow what is it? Show your work.**

* Banana. It can be shown using Naïve Bayes Method.



1. **Using a data set that relates stopping distance and speed, the following is the result of conducting linear regression on the results:**

**Provide an interpretation of the results. If I have a speed of 12 miles per hour, what will my stopping distance be?**

Residuals:

Min 1Q Median 3Q Max

-29.069 -9.525 -2.272 9.215 43.201

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -17.5791 6.7584 -2.601 0.0123 \*

speed 3.9324 0.4155 9.464 1.49e-12 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 15.38 on 48 degrees of freedom

Multiple R-squared: 0.6511, Adjusted R-squared: 0.6438

F-statistic: 89.57 on 1 and 48 DF, p-value: 1.49e-12

* **Residuals:** Residuals are the difference between actual/observed value and the predicted value on the regression line. In this scenario, it is the stopping distance. The equation shows that the model had different prediction value compared to the actual value.
* **Coefficients:** They are the estimates of the equation which tells us if there is a positive or negative correlation between Independent and Dependent variables. Coefficients have factors such as Estimates, Standard Error, T-Value and Pr(>t).
  + **Estimates:** If a car takes -17.5781 feet to stop (intercept), With every mile per hour increase in speed the stopping distance will increase by 3.9324 feet.
  + **Standard Error:** It is the standard error of the estimate which represents the average distance of the actual value from the regression line. Lower the number, better the estimate.
  + **T-Value:** T- Value measures the number of standard deviations that the estimate value is away from 0. Longer the distance from 0, null hypothesis can be rejected which shows more confidence in our model.
  + **Pr(>t):** Standard rule of P-value says that if P-value is less than 5% it boosts the confidence in data and proves significant relationship in the data.
* **Residual Standard Error:** Formula says that it is the square root of residual sum of squares divided by residual degrees of freedom. It shows if the model is suitable for fitting. Values close to 0 indicate that it is a better fit. 15.38 feet in this model.
* **R-Squared:** Represents the proportion of variance between dependent and independent variables. It measures if the actual data fits well in the model. The model shows around 65% R2 which we can infer as the variation in distance required to stop is 65% explanatory because of speed.
* **F-Statistic:** It is the determining factor that explains if there is a relationship between response variables and predictor variables. If value is more than 1, there is a strong relationship. This model has 89.57.

With the use of formula Y = mX + C, where Y= Stopping Distance, X=Speed, m = coefficient C = intercept. X= 12, m = 3.9324, C= -17.5791.

Y = 3.9324\*12 + (-17.5791)

**Y = 29.6097 Feet**

1. **Which variable selection methods are most likely computationally intensive – they will take longer to run than the other methods due to the amount of computation required or the amount of memory required? Explain.**

* The most Computationally Intensive methods are Forward Selection, Backward elimination, Stepwise regression, LASSO Regression, RIDGE Regression, KNN and Decision Trees.
* Forward Selection, Backward Elimination and Stepwise Regression methods have a process to generate a series of outcomes repetitively.
* LASSO and RIDGE methods are a combination iterative (repetitive process) and Simple Filter Methods.
* K-NN Test must find its nearest neighbor to match and find an outcome hence it is intensive.
* Decision Trees like Boosted, Pruned Regression and Classification methods are lengthy and intensive as they involve splitting data multiple times to check out different combinations.