

Group A

[Answer all the questions]

1. Answer any FIVE

5*1=5

- a) Define data.
- b) What is categorical variable?
- c) Define population and sample.
- d) What is selection bias?
- e) What is an irreducible error?
- f) What is a scatter plot?
- g) What is 95% confidence interval for linear regression?

2. Answer any FOUR

4*2.5=10

- a) Write the Data Science Process.
- b) What is Bias? How bias can happen in samples?
- c) Differentiate between regression and classification.
- d) Assume there are 4 features (X_1, X_2, X_3, X_4) in a data set, where the interaction of X_2 and X_4 might have some impact on the response. Write the appropriate Linear Regression model for it.
- e) What is a pie chart? Assume there are 100 items: 50 are red, 25 are green and rest are blue. Draw a pie chart for the data.
- f) Differentiate between linear regression and logistic regression.

3. Answer any TWO

2*5=10

- a) Perform KNN regression for the following data and show the predictions for all the X values. Assume $k = 2$.

X	1	2	3	4	5
Y	1	2	5	4	5

- b) What is F statistics? The true values and predicted values of a model are given below. Calculate the F statistics for the data and interpret the result.

X	1	2	3	4	5
Y	6	7	4	3	2
\hat{Y}	6.8	5.6	4.4	3.2	2

- c) What is MSE? Calculate the MSE for the above model in q(3b).

Group B

[Answer all the questions]

4. Answer any FIVE

5*1=5

- a) Define Akaike's Information Criterion.
- b) What are interacting predictors?
- c) What is overfitting?
- d) What is cross validation?
- e) What is Model Fitness, R^2 ?
- f) What is bootstrapping?
- g) Write a polynomial regression model of degree M.

270704
 1813715421

5. Answer any FOUR

4*2.5=10

- a) How regression can be done by decision trees?
- b) What is k-fold cross validation?
- c) What is LASSO Regression?
- d) Define Entropy and Information Gain.
- e) Compare between mean and median.
- f) Differentiate between parametric and non-parametric models with example.

6. Answer any TWO

2*5=10

- a) Write the steps of computing principal components of N data with J features.
- b) For the given data, approximate the values of B_0 and B_1 and write the equation for simple linear regression.

X	1	2	3	4	5
Y	6	7	4	3	2

- c) What is hypothesis testing? Write the steps of testing a hypothesis.

3(a)
 3(b)
 T-2(5)

Shahjalal University of Science and Technology
Department of Computer Science and Engineering
3rd Year 2nd Semester Final Examination' Dec 2019 (Session: 2016-17)
Course Code: CSE 345 Credits: 2 Course Title: Data Science
Time: 2 hrs Total Marks: 50

Group A

[Answer all the questions]

5x1=5

1. **Answer any FIVE**

- a) What is Structured Data?
- b) Define Quantitative Variable with examples.
- c) What is a Scatter Plot?
- d) Define Variance of a Sample.
- e) What is a Statistical Model?
- f) What is a Loss Function?
- g) How to interpret a P-Value?
- h) If B is a Linear Regression Coefficient and the Standard Error (B) = 1.5, then find the 95% Confidence Interval of B .

4x2.5=10

2. **Answer any FOUR**

- a) Describe the process of Data Science.
- b) Define Data, Distribution of Data, Population, Sample and Bias.
- c) Name and define different Measures of Centrality; Mention the Computational Complexity of each.
- d) What are the principles of Data Visualization?
- e) What is a Histogram? Draw the Histogram of the following data:

Height(Feet)	Frequency
0-2	0
2-4	1
4-5	4
5-6	8
6-8	2

- f) What are the Two main types of Prediction Problems we discussed in this course? Differentiate between them.

2x5=10

3. **Answer any TWO**

- a) Assume there are 4 predictors (X_1, X_2, X_3, X_4) in a data set, where the interaction of X_2 and X_4 might have some impact on the response.
 - i. Write the appropriate Linear Regression Model for it.
 - ii. Write the Final Model if you get the following P-Values after performing hypotheses testing on the significance of the predictors.

Coefficients	P-Value
B_0	0.00
B_1	0.09
B_2	0.00
B_3	0.03
B_4	0.00
B_5	1.00

- iii. What can you tell about the signification of the predictors?

- b) What is F statistics? The true values and predicted values of a model are given below. Calculate the F statistics for the data. And interpret the result.

X	1	2	3	4	5
Y	1	2	5	4	5
Y(Predicted)	3.5	3	3	5	4.5

- c) Perform Linear Regression on the following data to find the equation of the Regression line.

Data	Age X	Glucose Level Y
1	43	99
2	21	65
3	25	79
4	42	75

Group B

[Answer all the questions]

4. Answer any FIVE

5x1=5

- a) What is Logistic Regression?
- b) What is Ensemble Learning?
- c) What is a Random Forest?
- d) Define Entropy and Information Gain.
- e) Write the Bayes' Theorem.
- f) What is a ROC Curve?
- g) What is bootstrapping?
- h) Write a polynomial regression model of degree M.

5. Answer any FOUR

4x2.5=10

- a) What is Over-fitting? What are the causes of Over-fitting?
- b) Define the steps to choose the subset of significant predictors using K-fold cross validation.
- c) Why Regularization is used? Define L_1 and L_2 Regularization.
- d) Differentiate between Parametric and Non-Parametric Models with example.
- e) What is the most commonly used Loss Function? Define it and calculate its value from the following prediction.

X	1	2	3	4	5
Y	1	2	5	4	5
Y(Predicted)	3.5	3	3	5	4.5

- f) How to deal with Missing Values in data?

6. Answer any TWO

2x5=10

- a) What is Principal Component Analysis? Write the steps of computing principal components of N data with J features.
- b) Suppose you have written a classifier to detect which images in your favorite social network are selfies. You have tested your classifier with some data and got the following predictions.

	Target	Prediction
1	Selfie	Not Selfie
2	Selfie	Selfie
3	Not Selfie	Not Selfie
4	Selfie	Selfie
5	Not Selfie	Selfie
6	Not Selfie	Not Selfie
7	Not Selfie	Not Selfie
8	Selfie	Selfie
9	Selfie	Not Selfie
10	Not Selfie	Not Selfie

Compute the Confusion Matrix for above and calculate Accuracy from it.

- c) Briefly describe the process of building a Decision Tree.

Shahjalal University of Science and Technology

Department of Computer Science and Engineering

3rd year 2nd Semester Final Examination—December 2020 (Session 2017-18)

Course No.—**CSE 345**

Course Title—**Data Science**

Time—**5 Hours**

Credit: **3.00**

Total Marks#**30**

(Answer All the Questions)

Group A

- Determine the following **Five** statements as True or False. If false, write the correct verdict. $5 \times 1 = 5$
 - Tabular format is the most suitable representation of Data.
 - A Scatter Plot displays groups of numerical data through their quartiles.
 - A large variance in data indicates that the values are far from the mean.
 - Data visualization helps us to analyze and explore the data.
 - A statistical model is any algorithm that estimates the underlying function that represents the relationship between dependent and independent variables.
- Answer the following **Two** Questions. $2 \times 2.5 = 5$
 - Define Data, Population, Sample and Bias as briefly as possible.
 - If **B** is a Linear Regression Coefficient where **B** = 5 and the Standard Error (**B**) = 1.5, then find the 95% Confidence Interval of **B**.
- Consider the following dataset.

Day	Weather	Temperature	Wind	Play
1	Sunny	Hot	Strong	No
2	Cloudy	Mild	Weak	Yes
3	Sunny	Mild	Weak	Yes
4	Sunny	Mild	Strong	No
5	Rainy	Cool	Weak	No
6	Cloudy	Cool	Weak	Yes
7	Cloudy	Hot	Strong	No

Now, build a **decision tree** according to that dataset which will predict/decide whether you should play or not on a day given the *weather*, *temperature* and *wind* information of that day.

5

Group B

- Determine the following **Five** statements as True or False. If false, write the correct verdict. $5 \times 1 = 5$
 - Logistic Regression is supervised learning.
 - Ensemble methods use a single model to obtain better prediction.
 - A Random Forest Classifier takes votes from multiple Decision Trees.
 - A model that has a lower AIC (or BIC) is better than other models.
 - A ROC Curve illustrates the trade-off for all possible thresholds chosen for the two types of classification error.
- Answer the following **Two** Questions. $2 \times 2.5 = 5$
 - What is Cross Validation? How to perform K-fold Cross Validation?
 - What are the differences between Parametric and Non-Parametric Models?
- Define Entropy and Information Gain. Suppose you have tossed a 4 faced dice 1000 times where 1, 2, 3 and 4 showed up 250, 500, 125, and 125 times respectively. Calculate the Entropy for this dice.

5

Shahjalal University of Science and Technology

Department of Computer Science and Engineering

3rd year 2nd Semester Final Examination—December 2020 (Session 2017-18)

Course No.—CSE 345

Course Title—Data Science

Time—5 Hours

Credit: 3.00

Total Marks#30

(Answer All the Questions)

Group A

- Determine the following **Five** statements as True or False. If false, write the correct verdict. $5 \times 1 = 5$
 - For Quantitative Variables there is no inherent order among the values.
 - A Scatter Plot displays groups of numerical data through their quartiles.
 - Data visualization helps us to analyze and explore the data.
 - “Some samples are more likely to be selected”—this phenomenon is called volunteer bias
 - KNN Regression is a parametric model.
- Answer the following **Two** Questions. $2 \times 2.5 = 5$
 - Draw a Pie Chart with the following data:

Height (Feet)	Frequency
0-2	20
2-4	100
4-5	400
5-6	480

- Write down the differences between Regression and Classification.
- Write down the Linear Regression model for the following data, and find the equation of the Regression line. 5

Data	X	Y
1	4	10
2	2	6
3	3	8

Group B

- Determine the following **Five** statements as True or False. If false, write the correct verdict. $5 \times 1 = 5$
 - Logistic Regression is supervised learning.
 - Ensemble methods use a single model to obtain better prediction.
 - A Random Forest Classifier takes votes from multiple Decision Trees.
 - A model that has a lower AIC (or BIC) is better than other models.
 - In hypothesis testing, if $F > 1$, then we accept the null hypothesis.
- Answer the following **Two** Questions. $2 \times 2.5 = 5$
 - What is Over-fitting? Write the causes of Over-fitting.
 - Name different approaches to impute missing values in a variable.
- Suppose you have written a classifier to separate the images of Roshogolla and Chomchom. You have tested your classifier with some images and got the following predictions. Now make the Confusion matrix for that and calculate accuracy from it. 5

CSE 345: Introduction to Data Science

Time: 35 minutes, Marks: 20

- ✓ How do you measure node impurity in Decision Tree algorithm with Gini Indexing, Entropy and Information Gain? Discuss with an example. 8
- ✓ Data preparation is quite important in data science. What are the steps of data preparation? How do you handle missing values? 4
- ✓ Why should we need to address overfitting? How to address overfitting in pre-pruning and post-pruning? 8

Class Test#02

Time: 30 minutes, Marks: 20

✓ 1 Cluster the following eight points (with (x, y) representing locations) into three clusters: 8

A1(2, 10), A2(2, 5), A3(8, 4), A4(5, 8), A5(7, 5), A6(6, 4), A7(1, 2), A8(4, 9)

Initial cluster centers are: A1(2, 10), A4(5, 8) and A7(1, 2).

The distance function between two points $a = (x_1, y_1)$ and $b = (x_2, y_2)$ is defined as-

$$P(a, b) = |x_2 - x_1| + |y_2 - y_1|$$

Use K-Means Algorithm to find the three cluster centers after the second iteration.

- ✓ 2 What is Naïve Bayes algorithm? What advantages do we have of using this algorithm. 4
- 3 What is Bayesian network? Where can we apply this? 4
- 4 What are the differences between KNN and K-mean? 4

TT #02

Course: Machine Learning (SWE 427) (QT-A)

Marks: 20 (8+12) Time: 30 mins

1. What are support vectors? How do support vectors help to find the optimal margin of a model?
2. Given the data in the table below, reduce the dimensions from 2 to 1 using the PCA algorithm.

X_1	X_2
1	2
3	4
5	6
7	9

TT#01 Course: Machine Learning (SWE 427) (QT-A)
Marks: 20 Time: **30 mins**

1. Does gradient descent require a convex cost function to converge? Can we use Mean Squared Error for calculating gradient descent of Logistic Regression to converge to the global optima?
If not, why? **05**
2. Explain the role of the learning rate in gradient descent. What are the potential consequences of setting it too high or too low? **05**
3. Create a fictional case study where the improper use of regularization leads to significant model failures. What lessons can be learned from this scenario? **05**
4. How do filters extract features and how does pooling simplify them in a CNN? Explain in brief. **05**

TT#01 Course: Machine Learning (SWE 427) (QT-B)
Marks: 20 Time: **30 mins**

1. How does logistic regression differ from linear regression in terms of the nature of the dependent variable and the type of problems it solves? **05**
2. Explain the role of the learning rate in gradient descent. What are the potential consequences of setting it too high or too low? **05**
3. Create a fictional case study where the improper use of regularization leads to significant model failures. What lessons can be learned from this scenario? **05**
4. Provide a detailed mathematical breakdown of the forward propagation process in a simple neural network. **05**

TT#01 Course: Machine Learning (SWE 427) (QT-C)
Marks: 20 Time: **30 mins**

1. How can you visualize the decision boundary for a simple linear classifier? Discuss in brief. **05**
2. Explain the role of the learning rate in gradient descent. What are the potential consequences of setting it too high or too low? **05**
3. Create a fictional case study where the improper use of regularization leads to significant model failures. What lessons can be learned from this scenario? **05**
4. Evaluate the effectiveness of CNNs in image classification tasks. What problem it solved in deep learning. **05**

Shahjalal University of Science and Technology
Institute of Information and Communication Technology
Software Engineering
3rd Year 2nd Semester Final Examination' Dec 2019 (Session: 2016-17)
Course Code: SWE 335 Credits: 2 Course Title: Data Science
Time: 2 hrs Total Marks: 50

Group A
[Answer all the questions]

5*1=5

1. **Answer any FIVE**
- a) Define datum and data.
 - b) What is Residual Standard Error?
 - c) Define population and sample.
 - d) What is Messy data? What are the common causes of messiness?
 - e) What is selection bias?
 - f) State the complexities of computing mean and median of data.
 - g) What is 95% confidence interval for linear regression?

4*2.5=10

2. **Answer any FOUR**
- a) Write the Data Science Process.
 - b) How is data represented and stored? Give examples.
 - c) Differentiate between regression and classification.
 - d) Assume there are 4 features (X_1, X_2, X_3, X_4) in a data set, where the interaction of X_2 and X_4 might have some impact on the response. Write the appropriate Linear Regression model for it.
 - e) Write down the imputation methods for missing data.
 - f) What is cross validation? Describe k-fold cross validation.

2*5=10

3. **Answer any TWO**
- a) Perform KNN regression for the following data and show the predictions for all the X values. Assume $k = 2$.

X	1	2	3	4	5
Y	1	2	5	4	5

u. u5

- b) Use K-means to cluster the given data: {20, 3, 9, 10, 9, 3, 1, 8, 5, 3, 24, 2, 14, 7, 8, 23, 6, 12, 18} into 3 groups (use 2 iterations).
- c) What is MSE? Calculate the MSE for the given model.

X	1	2	3	4	5
Y	6	7	4	3	2
Y	6.8	5.6	4.4	3.2	2

3.625

Group B
[Answer all the questions]

5*1=5

Answer any FIVE

4. a) Write down the type of missingness.
- b) What are interacting predictors?
- c) What is overfitting?
- d) What is Variance and Standard Deviation?
- e) What is agglomerative clustering? Give example.
- f) Write a polynomial regression model of degree M.
- g) What is Web Scrapping?

4*2.5=10

Answer any FOUR

- a) What are the causes of over-fitting?
- b) Show the difference between Prediction and Estimation.
- c) What is LASSO Regression?
- d) Define Entropy and Information Gain.
- e) Use hierarchical cluster to cluster the given data {3, 7, 10, 16, 18, 20}. Show each steps.
- f) Differentiate between parametric and non-parametric models with example.

2*5=10

Answer any TWO

- a) For the given data, approximate the values of B_0 and B_1 and write the equation for simple linear regression.

X	1	2	3	4	5
Y	6	7	4	3	2

- b) Construct the decision tree to decide what to do in the evening, find the root (use the given dataset).

Deadline?	Is there a Party?	Lazy?	Activity
Urgent \	Yes	Yes	Party ✓
Urgent \	No	Yes	Study ✓
Near . ✓	Yes	Yes	Party ✓
None ~	Yes	No	Party ✓
None ~	No	Yes	Pub •
None ~	Yes	No	Party ✓
Near ✓	No	No	Study ✓
Near ✓	No	Yes	TV +
Near ✓	Yes	Yes	Party ✓
Urgent ✓	No	No	Study ✓

Party
Study
Pub
TV

- c) What is hypothesis testing? Write the steps of testing a hypothesis.

Group A

[Answer all the questions]

5x1=5

1. Answer any FIVE

- a) What is Structured Data?
- b) Define Quantitative Variable with examples.
- c) What is a Scatter Plot?
- d) Define Variance of a Sample.
- e) What is a Statistical Model?
- f) What is a Loss Function?
- g) How to interpret a P-Value?
- h) If B is a Linear Regression Coefficient and the Standard Error (B) = 1.5, then find the 95% Confidence Interval of B .

4x2.5=10

2. Answer any FOUR

- a) Describe the process of Data Science.
- b) Define Data, Distribution of Data, Population, Sample and Bias.
- c) Name and define different Measures of Centrality; Mention the Computational Complexity of each.
- d) What are the principles of Data Visualization?
- e) What is a Histogram? Draw the Histogram of the following data:

Height(Feet)	Frequency
0-2	0
2-4	1
4-5	4
5-6	8
6-8	2

- f) What are the Two main types of Prediction Problems we discussed in this course? Differentiate between them.

2x5=10

3. Answer any TWO

- a) Assume there are 4 predictors (X_1, X_2, X_3, X_4) in a data set, where the interaction of X_2 and X_4 might have some impact on the response.
 - i. Write the appropriate Linear Regression Model for it.
 - ii. Write the Final Model if you get the following P-Values after performing hypotheses testing on the significance of the predictors.

p value < 0.05
 $B_1 \rightarrow \text{sign}$

Coefficients	P-Value
B_0	0.00
B_1	0.09
B_2	0.00
B_3	0.03
B_4	0.00
B_5	1.00

$$\hat{Y} = B_0 + B_1 X_1 + B_2 X_2 + B_3 X_3 + B_4 X_4$$

- iii. What can you tell about the signification of the predictors?

- b) What is F statistics? The true values and predicted values of a model are given below. Calculate the F statistics for the data. And interpret the result.

X	1	2	3	4	5
Y	1	2	.5	4	5
Y(Predicted)	3.5	3	3	5	4.5

- c) Perform Linear Regression on the following data to find the equation of the Regression line.

Data	Age X	Glucose Level Y
1	43	99
2	21	65
3	25	79
4	42	75

Group B

[Answer all the questions]

5x1=5

4. **Answer any FIVE**
- a) What is Logistic Regression?
 - b) What is Ensemble Learning?
 - c) What is a Random Forest?
 - d) Define Entropy and Information Gain.
 - e) Write the Bayes' Theorem.
 - f) What is a ROC Curve?
 - g) What is bootstrapping?
 - h) Write a polynomial regression model of degree M.

4x2.5=10

5. **Answer any FOUR**
- a) What is Over-fitting? What are the causes of Over-fitting?
 - b) Define the steps to choose the subset of significant predictors using K-fold cross validation.
 - c) Why Regularization is used? Define L_1 and L_2 Regularization.
 - d) Differentiate between Parametric and Non-Parametric Models with example.
 - e) What is the most commonly used Loss Function? Define it and calculate its value from the following prediction.

X	1	2	3	4	5
Y	1	2	5	4	5
Y(Predicted)	3.5	3	3	5	4.5

- f) How to deal with Missing Values in data?

2x5=10

6. **Answer any TWO**
- a) What is Principal Component Analysis? Write the steps of computing principal components of N data with J features.
 - b) Suppose you have written a classifier to detect which images in your favorite social network are selfies. You have tested your classifier with some data and got the following predictions.

	Target	Prediction
1	Selfie	Not Selfie
2	Selfie	Selfie
3	Not Selfie	Not Selfie
4	Selfie	Selfie
5	Not Selfie	Selfie
6	Not Selfie	Not Selfie
7	Not Selfie	Not Selfie
8	Selfie	Selfie
9	Selfie	Not Selfie
10	Not Selfie	Not Selfie

Compute the Confusion Matrix for above and calculate Accuracy from it.

- c) Briefly describe the process of building a Decision Tree.

SWE 335 Term Test 1

*Introduction to
data science*

Marks: 25

Time: 40 Min

Assume this dataset below is about the unit price of a fabric. There are 10 shades of colors and 10 levels of qualities. Using this, answer the following questions.

	Color	Quality	Price
1	7	5	65
2	3	7	38
3	5	8	51
4	8	1	38
5	9	3	55
6	5	4	43
7	4	0	25
8	2	6	33
9	8	7	71
10	6	4	51

Results of Multiple Regression	
n	10
k	2
R-Square	0.850694
F	22.79061
p-value	0.000497

$$RSE = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

~~RSE~~ ?

1. Write FIVE questions that can be asked about this fabric.
2. Draw a scatter plot with Quality and Price and write the findings from it.
3. Interpret the regression results given in the above table. What is the hypothesis here? Do you accept it?
4. Assume Price is the response variable here. Write the equations for all possible linear regression models using this data up to 2nd order polynomials.
5. What is cross validation? Write how a 5-fold cross validation be done with the dataset to find the best model? What is your evaluation criterion?

Group A

[Answer all the questions]

1. Answer any FIVE

5x1=5

- a) What is Data Science?
- b) What is a Statistical Model?
- c) What is a Loss Function?
- d) What is Supervised Learning?
- e) What is Structured data? Give examples.
- f) What is a Confidence Interval?
- g) What is web scrapping?
- h) What is Bias-Variance Trade-off?

2. Answer any FOUR

4x2.5=10

- a) Assume there are two features (X_1, X_2) in a data set. Write the 3rd order Polynomial Regression model for it.
- b) Define Data, Distribution, Population, Sample and Bias.
- c) Write down the purpose(s) statistical modeling, with appropriate example(s).
- d) What are the things that we want to visualize about a data? Name the suitable plot(s) for each purpose.
- e) What are the things to consider while evaluating a model?
- f) What is Overfitting? Write why overfitting happens.

3. Answer any TWO

2x5=10
2

- a) i. Perform Linear Regression on the following dataset:

Height (inch)	Weight (lbs)
60	140
62	155
67	160
75	200

- ii. Re-estimate the responses using your model.
- iii. Calculate Loss and Fitness of the model.
- b) i. What is Hypothesis testing? Why do we do this?
- ii. Assume a model $Y = B_0 + B_1 X_1 + B_2 X_2 + B_3 X_3 + B_4 X_1 X_3 + \epsilon$. What is the final model if you get the following P-Values after performing hypotheses testing on the significance of the predictors?

Coefficients	P-Value
B_0	0.00
B_1	0.08
B_2	0.03
B_3	0.01
B_4	1.00

- iii. Write which predictor(s) are **not** significant.
- a) i. Write all possible regression models for the following data if you consider up to 2nd order polynomials and/or an interaction between predictors:

Y	X ₁	X ₂
140	60	22
155	62	25
179	70	20
192	71	15
200	72	14
215	78	11

- ii. What is Cross Validation? Write how K-fold cross validation can be used to select a suitable model.

Group B
[Answer all the questions]

4. Answer any FIVE

5x1=5

- a) What is Logistic Regression?
- b) What is Regularization?
- c) Define Bayes' Information Criterion.
- d) What is an ROC curve?
- e) Define Entropy.
- f) What is a Random Forest?
- g) Define Eigen Value and Eigen Vector.
- h) How categorical variables are used in computation?

5. Answer any FOUR

4x2.5=10

- a) Differentiate between regression, classification and clustering.
- b) What are the differences between Parametric and Non-Parametric Models?
- c) How regression can be done by decision trees?
- d) How to avoid Overfitting?
- e) Define Entropy. Suppose you have tossed a 4 faced dice 1000 times where 1, 2, 3 and 4 showed up 250, 500, 125, and 125 times respectively. Calculate the Entropy for this dice. What can you say about this dice?
- f) What is Imputation? How to impute missing values in data?

6. Answer any TWO

2x5=10

- i. What is Information Gain?
- ii. You want to buy a car and have the following models available. Build a Decision tree with the data, show the calculations.

	Age	Mileage	Road Tested	Buy
1	Recent	Low	Yes	Buy
2	Recent	High	Yes	Buy
3	Old	Low	No	Don't buy
4	Recent	High	No	Don't buy

- iii. What are the limitation(s) of a Decision Tree?

- b) Suppose you have written a classifier to label your pictures of sad faces and happy faces. You have tested your classifier with some images and got the following predictions:

	Target	Prediction
1	Happy	Sad
2	Happy	Sad
3	Sad	Sad
4	Happy	Happy
5	Sad	Happy
6	Sad	Sad
7	Sad	Sad
8	Happy	Happy
9	Happy	Sad
10	Sad	Sad

- i. Make the Confusion matrix.
 - ii. Calculate Accuracy, Sensitivity, Precision and Recall of your classifier.
- c) i. What is Principal Component Analysis?
 - ii. Write the steps of computing principal components of N data with J features.
 - iii. How can we get the original data back?