M.Sc. Examination, 2025
Semester-IV
Statistics
Course: MSC-41(Reliability Analysis)
Time: Three Hours Full Marks: 40

Questions are of value as indicated in the margin
Notations have their usual meanings

Answer **Question No. 1** and **any four** of the remaining questions

1. Answer any ***four*** of the following questions: 2x4=8
(a)Define k-out of- n: G system.
(b) Write down the UMVUE of reliability function, R(t) for exponentially distributed lifetime based on a random sample of size n.
(c) State the "Lack of Memory" property.
(d) What are irrelevant and redundant components in a system?
(e) When is a distribution said to be NBU?
(f) What do you understand by accelerated life test?

2. Briefly describe failure rate and reliability function. Establish the relationship between them. Show that, under suitable assumptions, failure rate of a series system is the sum of individual component failure rates. 2+3+3

3. Define Failure Rate Average. Prove that constancy of ratio of Failure Rate to Failure Rate Average is the characterization of the Weibull distribution. 2+6

4. (i) Explain Type-I and Type-II censorings in life testing.
(ii) Obtain the maximum likelihood estimator of the parameter of an exponential distribution using both type of censoring. Hence obtain the estimates of reliability. 3+ (4+1)

5. Derive a bivariate exponential distribution from a fatal shock model. Does this BVED satisfy the lack of memory property? Work out the regression functions. 2+2+4

6. (i) What do you understand by IFR and DFR in reliability study.
(ii) Show that in case of Weibull distribution the nature of the failure rate function depends on its shape parameter, and the distribution can be used in all three phases of bathtub curve.
(iii) For a 2-out-of-3 system, let the reliability of each component be 'p'. Find the reliability of the system assuming the components are statistically independent. Draw a rough sketch of the system reliability in relation to p? 2+3+3

7. (i) Let $T_1, T_2, ..., T_n$ be independently and identically distributed random variables having reliability function given by

$$R(t) = 1 - \theta t + o(t) \text{ as } t \to 0.$$

Show that $X_n = n.\min\left(T_1, T_2, ..., T_n\right)$ has asymptotically an exponential distribution as $n \to \infty$.
(ii) Define Mean Residual Life. Discuss how a life distribution is classified based on mean residual life. 4+4

M.Sc. Examination 2025
Semester-IV
Statistics
Course: MSC-42 (Bayesian Inference)

Time: 3 Hours **Full Marks: 40**

Questions are of value as indicated in the margin
Notations have their usual meanings

Answer **any four** questions

1. a) Define exchangeability. Show that if $y_1,\cdots,y_n$ are conditionally i.i.d for given θ, then marginally $y_1,\cdots,y_n$ are exchangeable. State de Finetti's theorem in this regard.
   b) Let $\theta \sim Gamma(a,b)$ and $y_1,\cdots,y_n|\theta \sim iid\ Poisson(\theta)$. Obtain the posterior distribution of $\theta|y_1,\cdots,y_n$. For a new observation $\tilde{y}$, find the expression of the predictive mean and variance under (a). 5+5

2. (a) Let θ be the probability of success in a Binomial model with n observations. Consider the following prior distribution for θ:

$$p(\theta) = \frac{3}{4}Beta(a_1,b_1) + \frac{1}{4}Beta(a_2,b_2)$$

Find the posterior distribution and show that it is a mixture of two distributions you know. Identify these distributions.

b) Let $Y \sim Bin(n,\theta)$. Obtain Jeffrey's prior distribution $p_J(\theta)$ for this model. 5+5

3. (a) Let $x_1,\cdots,x_n$ be i.i.d $N(\mu,\tau^2)$. Assume that $\mu \sim N(\mu_0,\tau_0^2)$ apriori. Compute the posterior distribution of μ_n assuming τ^2 known.
   (b) Show that for Bernoulli (θ), under conjugate prior set up, the posterior expectation of θ is a convex combination of the prior expectation and the sample average. Discuss the case when sample size $n \to \infty$. 5+5

4. (a) What is Gibbs sampling? Describe general properties of a Gibbs sampler.
   (b) For a normal set up with semi-conjugate prior for the variance and normal prior for mean, describe the Gibbs sampling process in details. 5+5

5. (a) Describe a Bayesian method to compare two groups.
   (b) How will you generalize this method for multiple groups? Describe the process with the help of the hierarchical normal model. 5+5

6. (a) Describe the role of Markov chain Monte Carlo (MCMC) simulation in Bayesian analysis. How does it differ from Monte Carlo (MC) method? Give an example.
   (b) How does the correlation of the MCMC samples affect posterior approximation?
   (c) Show that, in general,

$$Var_{MCMC} \geq Var_{MC}$$

5+2+3

M.Sc. Semester IV Examination 2025
Subject: Statistics
Paper: MSC 43/MSS 09
(Introductory Data Science and Statistical Machine Learning)

Full Marks: 40 Time: 3 hours

Answer any **four** of the following six questions of equal marks.
(Notations carry usual meanings)

1. (a) What do you mean by Data Science? Why do you think this is important in the present time?
   (b) Describe different components of Data Science with examples.

   4+6

2. (a) What are outliers? How does it impact data-driven decisions? Explain methods of identifying outliers in a dataset.
   (b) Explain Data Scaling and its necessity. Describe different types of Data Scaling methods.

   5+5

3. (a) What do you understand by Unsupervised Learning? Describe its types with examples.
   (b) Differentiate between Supervised and unsupervised Learning. Give a practical example scenario for each.

   5+5

4. (a) Explain how you would proceed to predict the outcome of a binary dependent variable. Provide a mathematical foundation and interpret the coefficients of the model.
   (b) Also provide how you would measure the model performance for deployment in operation.

   7+3

5. (a) What is predictive text analysis? Why is it needed? Describe different steps in predictive text analysis with an example.
   (b) What is association rule mining? Explain the terms support, confidence, and lift in association rule mining. How do these metrics help evaluate a rule?

   5+5

6. (a) What is topic modeling, and when is it used? Define different terms used in topic modeling and demonstrate the concept with an intuitive example.

   (b) What is the role of pruning in the Apriori algorithm? How does it help to improve efficiency?

6+4

M.Sc Semester IV Examination, 2025

Statistics

MSC-44(Practical)

Time: Four Hours   Full Marks: 40

(One may use a Computer, if necessary)

1. Plot the failure rate function of (1) series system, (2) parallel system, each consisting of two independent components, the first having a failure time distribution with i.e.

$$f_1(x) = 4.exp\,(-4x)$$

and the second having the i.e.

$$f_2(x) = 2.7.x^2.exp\,(-0.9x^3)$$

(take at least 6 points).

Also, find the Reliability of each system for a mission time of 100 units. 6

2. 100 electronic tubes of a certain type were tested. The test terminated after the first 15 tubes blew. Time failures occur after the following hours:

40,60,90,120,195,260,350,420,501,626,650,730,815,910,970.

Estimate R(t) at t=400 hrs. 5

3. Consider a document containing the sentence:

”data science is about data and science”

(a) Compute the Term Frequency (TF) for the words ”data” and ”science”.

(b) Calculate the proportion (in percentage) of these two words in the document. 1+1

4. Given a transactional dataset with 5 transactions:

T1 : {Bread, Milk}
T2 : {Bread, Diaper, Fruits, Eggs}
T3 : {Milk, Diaper, Fruits, Cola}
T4 : {Bread, Milk, Diaper, Fruits}
T5 : {Bread, Milk, Diaper, Cola}

(a) Calculate the support for the itemset {Bread, Milk}.

(b) Calculate the confidence for the rule {Bread} → {Milk}.

From the same transactional data (given above in Problem 4), calculate the lift for the rule {Diaper} → {Fruits}.

$$Lift(A \rightarrow B) = \frac{Confidence(A \rightarrow B)}{Support\;(B)}$$

Clearly show each calculation step and interpret the results of the obtained values. 2+2+2

5. Consider the following dataset:

ID	Humidity	Play Tennis?
1	High	No
2	High	No
3	Normal	Yes
4	Normal	Yes
5	Normal	No
6	High	No
7	Normal	Yes
8	High	Yes

(a) Calculate the entropy of the entire dataset with respect to the ”Play Tennis”decision.
(b) Calculate the entropy of each subset created by splitting on ”Humidity”.
(c) Compute the Information Gain obtained by splitting on ”Humidity” and interpret your result. 1+2+2

6. Generate 100 samples from a Bivariate Normal((last digit of your roll no.),(last but one digit of your roll number), 2,2,0.8) distribution using Gibbs sampling. 6

7. Last year, one of your friends, Pritha, got fed up with the number of fraud risk phone calls she was receiving. She set out with a goal of modeling the rate λ, typical number of fraud risk calls received per day. Prior to collecting any data, her guess was that this rate was most likely around 5 calls per day, but could also reasonably range between 2 and 7 calls per day. To learn more, she planned to record the number of fraud risk phone calls on each of 4 sampled days (Y_1, Y_2, Y_3, Y_4). The data is found to be (6,2,2,1).

(a). Using an appropriate conjugate prior, find the posterior distribution of λ and plot it with the prior distribution on the same space.
(b). Find posterior mean and variance and 95% credible interval for λ. 5

8. Notebook & Viva-voce 5

M.Sc. Examination, 2024
Semester-IV
Statistics
Course: MSC-41(Reliability Analysis)
Time: Three Hours Full Marks: 40

Questions are of value as indicated in the margin
Notations have their usual meanings

Answer **Question No. 1** and **any four** of the remaining questions

1. Answer any ***four*** of the following questions: 2x4=8

(a)What do you mean by standby component of a system?

(b) Write down the UMVUE of reliability function, R(t) for exponentially distributed lifetime.

(c)What do you understand by accelerated life test?

(d) Define availability at time t and limiting average availability of a system.

(e) Differentiate between system reliability and software reliability.

(f) For a proportional failure time model, derive the reliability function in term of the baseline reliability function.

2. Briefly describe failure rate and reliability function. Establish the relationship between them. Show that, under suitable assumptions, failure rate of a series system is the sum of individual component failure rates. 2+3+3

3. Define Failure Rate Average. Prove that constancy of ratio of Failure Rate to Failure Rate Average is the characterization of the Weibull distribution. 2+6

4. (i) Explain Time censoring and Number censoring in life testing.

(ii) Obtain the maximum likelihood estimators of the scale and shape parameters of Weibull distribution using Number censoring. Hence obtain the estimate of reliability. 3+ (4+1)

5. Derive a bivariate exponential distribution from a fatal shock model. Does this BVED satisfy the lack of memory property? Work out the regression functions. 2+2+4

6.(i) Prove that hazard rate r(t) is increasing in t>0 iff $\bar{F}(t+x)/\bar{F}(t)$ is decreasing in t for each $x \geq 0$, where $\bar{F}(t) = 1 - F(t)$ and $F(t)$ is the distribution function of the random variable T.

(ii) When a distribution is said to belong to IFRA class? 5+3

7. (i) Let $T_1, T_2, ..., T_n$ be independently and identically distributed random variables having reliability function given by

$$R(t) = 1 - \lambda t + o(t) \text{ as } t \to 0.$$

Show that $X_n = n.\min(T_1, T_2, ..., T_n)$ has asymptotically an exponential distribution as $n \to \infty$.

(ii) Explain type I and type II censoring used in life testing. Estimate the mean parameter of an exponential distribution using both types of censoring.

4+4

M.Sc. Examination 2024
Semester-IV
Statistics
Course: MSC-42 (Bayesian Inference)

Time: 3 Hours **Full Marks: 40**

Questions are of value as indicated in the margin
Notations have their usual meanings

Answer **any four** questions

1. (a) What is a conjugate prior?
   (b) Show that for Bernoulli (θ), under conjugate prior set up, the posterior expectation of θ is a convex combination of the prior expectation and the sample average. Discuss the case when sample size $n \to \infty$.
   (c) How will you modify your results if uniform prior is considered for θ?
   2+5+3
2. (a) What is Gibbs sampling? Describe general properties of a Gibbs sampler.
   (b) For a normal set up with semi-conjugate prior for the variance and normal prior for mean, describe the Gibbs sampling process in details. 5+5
3. (a) Let $\theta \sim Gamma(a, b)$ and $y_1, \cdots, y_n | \theta \sim Poisson(\theta)$. Obtain the posterior distribution of $\theta | y_1, \cdots, y_n$.
   (b) For a new observation $\tilde{y}$, find the expression of the predictive mean and variance under (a).
   6+4
4. (a) What is HPD region in Bayesian analysis?
   (b) Describe a Bayesian method to compare two groups.
   (c) Show that, in general,
   $Var_{MCMC} \geq Var_{MC}$ 2+5+3
5. (a) Describe the procedure for the Bayesian analysis of a regression model using semi-conjugate prior distributions.
   (b) What is "g-prior"? How will you modify the analysis if you are using "g-prior" in (a)?
   6+4
6. (a) What is Metropolis algorithm? How is it different from Metropolis-Hastings algorithm?
   (b) Show that Gibbs sampling is a special type of Metropolis-Hastings algorithm.
   6+4

M.Sc. Semester IV Examination 2024
Subject: Statistics
Paper: MSC 43/MSS 09
(Introductory Data Science and Statistical Machine Learning)

Full Marks: 40 Time: 3 hours

Answer any <u>four</u> of the following six questions of equal marks.
(Notations carry usual meanings)

1. (a) What are the common techniques used in data preprocessing, and why are they essential before model training?
   (b) Describe with examples how would you handle missing values in a data set to conduct a useful analysis and model buildings.

   6+4

2. (a) What do you understand by suvervised machine learning? Describe its different types with examples.
   (b) What is ensemble learning, and why is it needed? What are the methods of ensembling different models?

   5+5

3. (a) Explain 'Boosting' with an illustrative example.
   (b) What is gradient descent algorithm? When is it used? Explain it by an example.

   5+5

4. (a) Define the terms: i) Entropy, ii) Information Gain and iii) Gini impurity
   (b) Explain how those are used in classification problems.

   3+7

5. (a) Mention a few challenges faced in Text Analysis. Also describe some practical applications of text analysis.
   (b) Define Support Vector Machine (SVM) and mention its objective. In this connection also explain the terms: Margin, Support Vectors, Regularization Parameter and Soft Margin Classification.

   4+6

6. Compare different types of decision tree algorithms with respect to their features, limitations and uses.

   10

M.Sc Semester IV Examination, 2024
Statistics
MSC-44(Practical)

Time: Three Hours Full Marks: 40

(One may use Computer, if necessary)

1. 1. Plot the failure rate function of (1) series system (2) parallel system, each consisting of two independent components, the first having a failure time distribution with i.e. $f_1(x) = 3\,exp\,(-3x)$ and the second having the i.e. $f_2(x) = 2.4\,x^2 exp\,(-0.8x^3)$ (take at least 6 points).
 Also find the Reliability of each system for a mission time 100 units. 6

2. 100 electronic tubes of a certain type were tested. The test terminated after the first 15 tubes blew. Time failures occur after the following hours:
40,65,90,120,195,265,350,420,501,620,655,730,815,910,980.
Estimate R(t) at t=500 hrs. 5

3. Suppose it is known that for a particular drug, the success rate (θ) varies from 0.2 to 0.6. The drug was administered to 20 (n) healthy volunteers and we observed 15 (r) success. Using this information find a prior distribution for the success rate θ. Find the predictive posterior of the number of success in next 40 trials. Find an estimate of probability of getting atleast 25 successes out of 40 trials. 7

4. Consider the following data on the wing length in millimeters of nine members of a species of midge (small, two-winged flies).
1.64, 1.70, 1.72, 1.74, 1.82, 1.82, 1.82, 1.90, 2.08
From these nine measurements we wish to make inference on the population mean θ. Studies from other populations suggest that wing lengths are typically around 1.9 mm and standard deviation should not be too far from 0.01.
Find the posterior distribution of θ and 95% credible interval for θ. 5

5. Upload the supplied data file in your system and conduct a thorough data analysis to build a predictive model. The data set have six variables namely, 'Student_Id', 'Study_Hours', 'Assignment_Score', 'Class_Attendance' (in %), 'Participation', 'Final_Exam_Score'.The answer report must include the following:
a) Scrutinizing the data set and taking appropriate measures to handle data discrepancies.
b) Visual analytics and interpretations of 'Study_Hours' and 'Final_Exam_Score'.

c) Is there any significant difference in 'Final_Exam_Score' between students with different levels of Participation (High, Medium, Low)? Answer with reasons.

d) Develop a predicting model to classify students' performance as 'Pass' or 'Fail' based on a threshold of 50 in the 'Final_Exam_Score' using at least two methods. Use 'Study_Hours', 'Assignment_Scores', 'Class_Attendance' and 'Participation' as predictors.

Compare the model performances for prediction using the confusion matrices.

Associated R or Python code should be written on the script. 12

6. Notebook & Viva-voce 5

M.Sc. Examination, 2023
Semester-IV
Statistics
Course: MSC-41(Reliability Analysis)
Time: Three Hours Full Marks: 40

Questions are of value as indicated in the margin
Notations have their usual meanings

Answer **Question No. 1** and **any four** of the remaining questions

1. Answer any ***four*** of the following questions: 2x4=8
(a)What do you mean by standby component of a system?
(b) Write down the UMVUE of reliability function, R(t) for exponentially distributed lifetime.
(c) State the "Lack of Memory" property.
(d) What are irrelevant and redundant components in a system?
(e) When is a distribution said to be NBU?
(f) What do you understand by accelerated life test?

2. Briefly describe failure rate and reliability function. Establish the relationship between them. Show that, under suitable assumptions, failure rate of a series system is the sum of individual component failure rates. 2+3+3

3. Define Failure Rate Average. Prove that constancy of ratio of Failure Rate to Failure Rate Average is the characterization of the Weibull distribution. 2+6

4. (i) Explain Time censoring and Number censoring in life testing.
(ii) Obtain the maximum likelihood estimators of the scale and shape parameters of Weibull distribution using Number censoring. Hence obtain the estimate of reliability. 3+ (4+1)

5. Derive a bivariate exponential distribution from a fatal shock model. Does this BVED satisfy the lack of memory property? Work out the regression functions. 2+2+4

6. (i) Distinguish between IFR and DFR.
(ii) Show that in case of Weibull distribution failure rate depends on its shape parameter and the distribution can be used in all three phases of bathtub curve.
(iii) For a 2-out-of-3 system, let the reliability of each component be 'p'. Find the reliability of the system assuming the components are statistically independent. How does the system reliability behave in relation to p? 2+3+3

7. (i) Let $T_1, T_2, ..., T_n$ be independently and identically distributed random variables having reliability function given by

$$R(t) = 1 - \lambda t + o(t) \text{ as } t \to 0.$$

Show that $X_n = n.\min\left(T_1, T_2, ..., T_n\right)$ has asymptotically an exponential distribution as $n \to \infty$.
(ii) Define Mean Residual Life. Discuss how a life distribution is classified based on mean residual life. 4+4

M.Sc. Examination 2023
Semester-IV
Statistics
Course: MSC-42 (Bayesian Inference)

Time: 3 Hours **Full Marks: 40**

Questions are of value as indicated in the margin
Notations have their usual meanings

Answer **any four** questions

1. (a) What is a conjugate prior? Give an example.
(b)Show that for Poisson (θ), under conjugate prior set up, the posterior expectation of θ is a convex combination of the prior expectation and the sample average.
(c)For a new observation $\tilde{y}$, find the expression of the predictive mean and variance under (b).
2+5+3
2. (a) What is Gibbs sampling? Describe general properties of a Gibbs sampler.
(b)For a normal set up with semi-conjugate prior for the variance and normal prior for mean, describe the Gibbs sampling process in details. 5+5

3. Let $\boldsymbol{Y_1}, \boldsymbol{Y_2}, \ldots \boldsymbol{Y_n} \sim N(\boldsymbol{\theta}, \boldsymbol{\Sigma})$. Assuming normal prior for the mean vector $\boldsymbol{\theta}$ and Wishart prior for the dispersion matrix $\boldsymbol{\Sigma}$, find the posterior distributions of $\boldsymbol{\theta}|\boldsymbol{Y_1}, \boldsymbol{Y_2}, \ldots \boldsymbol{Y_n}$ and $\boldsymbol{\Sigma}|\boldsymbol{Y_1}, \boldsymbol{Y_2}, \ldots \boldsymbol{Y_n}$.
10
4. (a) Describe a Bayesian method to compare two groups.
(b) How will you generalize this method for multiple groups? Describe the process with the help of the hierarchical normal model. 5+5

5. (a) Describe the procedure for the Bayesian analysis of a regression model using semi-conjugate prior distributions.
(b) What is "g-prior"? How will you modify the analysis if you are using "g-prior" in (a)?
6+4
6. (a) Describe the role of Markov chain Monte Carlo (MCMC) in Bayesian analysis. How does it differ from Monte Carlo (MC) method? Give an example.
(b) How does the correlation of the MCMC samples affect posterior approximation?
(c) Show that, in general,
$$Var_{MCMC} \geq Var_{MC}$$
5+2+3

M.Sc. Semester IV Examination 2023

Subject: Statistics

Paper: MSC43/MSS 09

(Introductory Data Science and Statistical Machine Learning)

Time: Three Hours | Full Marks: 40

Answer any **four** of six questions.

1. a) What do you mean by feature scaling and centering? Differentiate between them with examples.
   b) Briefly describe the ways of detecting outliers in a dataset and how you would analyze a dataset that contains outliers.

5+5

2. a) What are the different advantages and disadvantages of the decision tree algorithm?
   b) Why should one prefer a (random) forest (collection of trees) to a single tree?

7+3

3. a) Explain why the linear regression technique is not applicable to model the target variable's yes/no or zero/one data. Describe a suitable technique to model such types of data with reasons.
   b) In comparison to other techniques what is/are the additional advantage(s) of the Logistic Regression technique in classification problems?

6+4

4. What is Kernel-trick in Support Vector Machines (SVM)? Why is it important? Answer with an illustration. Cite examples of a few Kernel functions. Is SVM sensitive to feature scaling? 10

5. a) What do you understand by unstructured data? Give examples.
   b) Explain the concept of text analysis with examples. How does it differ from Natural Language Processing (NLP)? In this context also explain the terms 'Tokenization', 'Lammentization' and 'Stop words.

3+7

[Turn Over]

6. a) Consider the following transactional data:

TID	Items
1	Bread, Milk
2	Bread, Diaper, Beer, Eggs
3	Milk, Diaper, Beer, Coke
4	Bread, Milk, Diaper, Beer
5	Bread, Milk, Diaper, Coke

Find the support of the item set {Milk, Diaper, Bread}. Take the rule {Milk, Diaper}→ {Beer} then find and interpret the Confidence and Lift of the rule.

b) Mention some advantages and disadvantages of Association Rule Mining.

6+4

M.Sc Semester IV Examination, 2023

Statistics

MSC-44(Practical)

Time: Four Hours   Full Marks: 40

(One may use Computer, if necessary)

1. In order to estimate the mean burning time of a particular brand of bulb, 30 bulbs were left burning. The bulbs that failed are not replaced upon failure. The following burning time (in hours) were recorded: 20, 27, 52, 61, 110, 122, 150, 214, 232, 238, 371, 393, 426, 445, 472, 503, 526, 581, 627, 698, 805, 909, 976, 1001, 1016, 1033, 1086, 1192, 1322, 1581.
   Calculate the maximum likelihood estimate and minimum variance unbiased estimate of R(t) at t=300 hrs. 6

2. 800 electronic components were placed on life tests. The system is observed at 3 hrs, 6 hrs, 9 hrs…, 30 hrs. The number of failures is noted. Calculate the hazard rates, and plot it in diagram and comment. 5

Failure data for 800 electronic components

Time interval in hours	Number of Failures in interval
0-3	190
3-6	72
6-9	46
9-12	30
12-15	17
15-18	13
18-21	14
21-24	9
24-27	6
27-30	3

3. Consider the following data on the weight of poultry on different days of measure:
Days of measure (X): 8, 15, 22, 29, 36, 43, 50
Weight in grams (Y): 177, 236, 285, 350, 376, 401, 430
The regression equation is:
$$y = \mu + \beta(x - \bar{x}) + e_{ij}$$
where $e_{ij} \sim N(0, \sigma^2)$
Perform a Bayesian linear regression of Y on X assuming informative and non-informative priors for the parameters involved in the above model. 6

4. Consider the following data on the wing length in millimeters of nine members of a species of midge (small, two-winged flies).
1.64, 1.70, 1.72, 1.74, 1.82, 1.82, 1.82, 1.90, 2.08
From these nine measurements we wish to make inference on the population mean θ. Studies from other populations suggest that wing lengths are typically around 1.9 mm and standard deviation should not be too far from 0.01.
Find the posterior distribution of θ and 95% credible interval for θ. Compare this credible interval with 95% confidence interval for θ and comment. 6

5. Classic Housing Bank wishes to automate its housing loan processing system. It has provided its previous applicants' data along with their loan status 'Yes' or 'No' i.e., approved or not approved. They consider most of the information (e.g., Gender, Education Level, Annual Income, Dependent members in the family, Loan amount, credit history, etc.) collected in the loan application form. Given their data, develop a prediction model for predicting an applicant's chance of getting the loan. Show the accuracy of your model in the appropriate form. Interpret your model.
Also, comment on the followings:
a. How does the loan approval chance vary if the applicant is a Male than a female?
b. How does the loan approval chance vary if the applicant has a bad credit history?
c. From the supplied new applicant's data what percentage of the applicants may get their loans approved?
(Data files will be provided during the exam)
Note: Credit History: 1 indicates the applicant has a credit history 12

6. Notebook & Viva-voce 5