

INSTITUTE OF ACTUARIES OF INDIA
EXAMINATIONS

May 2026

CS2B - Risk Modelling and Survival Analysis

Time allowed: 1 Hours 45 Minutes

Total Marks: 100

Q. 1) You are provided with a dataset, **employee_attrition.csv**, which contains the following columns:

- **Attrition:** Whether employee left the company (1 = Yes, 0 = No)
 - Type: Categorical (0/1)
 - 1 = Employee has left the company
 - 0 = Employee is still working
- **Age:**
 - Type: Numeric
 - Age of the employee
- **Department:**
 - Type: Categorical (e.g., HR, Sales, Tech etc.)
 - Department in which employee is working
- **MonthlyIncome:**
 - Type: Numeric
 - Salary earned per month (₹)
- **YearsAtCompany:**
 - Type: Numeric
 - Number of years employee has worked in the company
- **JobSatisfaction** (1–4 scale):
 - Type: Ordinal
 - 1 = Low Satisfaction
 - 4 = High Satisfaction
- **OverTime:**
 - Type: Categorical (Yes/No)
 - Whether employee works overtime regularly
- **NumCompaniesWorked:**
 - Type: Numeric
 - Number of companies employee has worked in before
- **EducationField:**
 - Type: Categorical (e.g., HR, Marketing, Engineering, etc.)
 - Field of education of employee

You are required to use the packages “dplyr”, “rpart”, “ggplot2”, “rpart.plot”, “scales”.

Using R, perform the following tasks:

- i) Import the dataset and recode categorical variables into factors, as needed. [2]
- ii) Check for missing values and replace them using Median value for the respective field/column. [2]

- iii) Examine the relationship between **Attrition** and other non-numerical variables (Overtime, Department, Education Field) using `table()` function to create appropriate summaries of count of unique values by clearly showing rows and columns. Also, Using R code, convert counts into row-wise proportion. [3]
- iv) Interpret the results of part (iii) and comment on whether overtime/Department/EducationField impacts employee attrition. [3]
- v) Create a well labelled boxplot image to analyse the relationship between attrition and numerical variables: **Age, MonthlyIncome, YearsAtCompany, NumCompaniesWorked, Job satisfaction** using R. [5]
- vi) Based on the visualizations, provide clear interpretations and key insights from the results. [3]
- vii) Build a decision tree model using variables OverTime + YearsAtCompany + JobSatisfaction and make inference from the output. [8]
- viii) Decision Tree model will give predicted probabilities. Obtain predicted probabilities of attrition from the model and paste the data showing predicted probabilities. Paste the first 20 rows of the data. [2]
- ix) Classify employees using cutoffs of **0.4 and 0.6** and paste the output. Paste the first 20 rows of the data. [3]
- x) Rank employees based on predicted attrition probability. Paste the first 20 rows of the data. [2]
- xi) Divide employees into **deciles (10 groups)** based on risk. Paste the first 20 rows of the data. [2]
- xii) Calculate **attrition rate in each decile**. [3]
- xiii) Plot deciles vs attrition rate and interpret the results. [3]

[41 Marks]

Q. 2) You are provided with a dataset, **hospital_stay.csv**, which contains the following columns:

- time: Days until discharge or death
- event: (1 = death, 0 = discharged)
- treatment_group: (A, B, C)
- age
- comorbidity_score
- icu_admission (1 = Yes, 0 = No)

You are required to use the packages “survival”, “survminer”, “dplyr”

Using R, answer the following questions: -

- i) Plot Kaplan-Meier curves for the entire dataset. [3]
 - ii) Plot Kaplan -Meier curves to compare across treatment groups on same graph. [3]
 - iii) Perform a log-rank test to test whether survival curves obtained in part (ii) above differ statistically and significantly across groups and interpret the result. [3]
 - iv) Check proportional hazards assumption before Cox model and interpret the results. [4]
 - v) Fit Cox model including interaction term “treatment_group $\times$ icu_admission”, age, comorbidity score and interpret the interaction effect. [5]
 - vi) Discuss limitations of Cox model in medical datasets. [3]
- [21 Marks]**

Q. 3) Before answering this question, run the following R code:

```
set.seed(1450)
x = arima.sim(list(order=c(1,0,1), ar=0.6, ma=-0.4), n=360)

fit = arima(x, order=c(1,0,1))
```

You are required to use the packages “forecast”, “dplyr”

Solve the following questions:

- i) Construct 95% confidence intervals for both the AR(1) and MA(1) parameters. [4]
- ii) Using the fitted model, obtain five-step-ahead forecasts of the time series. Display the predicted values. [3]
- iii) Create an R matrix B (5×3) containing:
 - a) forecast values
 - b) forecast standard errors
 - c) upper 95% prediction limits. [4]
- iv) Construct R code to produce a time series plot containing:
 - a) the observed data
 - b) the five-step-ahead forecasts
 - c) the corresponding 95% prediction intervals. [3]

[14 Marks]

Q. 4) A company models customer loyalty with states:

$$S = \{1 = \text{New}, 2 = \text{Active}, 3 = \text{Dormant}, 4 = \text{Lost}\}$$

Transition matrix is estimated from data.

Suppose you have observed states

```
current <- c(1,2,2,3,2,1,3,4,2,3)
```

```
Future <- c(2,2,3,4,3,2,4,4,1,4)
```

- i) Construct a transition matrix from given simulated data and convert it into probabilities. [4]
 - ii) Test whether state “Lost” (state 4) is absorbing. [2]
 - iii) Simulate Markov chain for 5,000 customers and estimate long-run proportions. Use seed 1 and compare it with the theoretical stationary distribution. (Candidates are NOT required to print the output of the simulation.) [5]
 - iv) Compute the n-step transition probabilities till 3 steps. [4]
 - v) Starting from state 1 (New), find the probability of being in state 4 after:
 - a) 2 steps
 - b) 5 steps [4]
 - vi) Given a customer is in state 2 (Active) at time t, what is the probability that they will be in state 4 (Lost) at time t+2 given they are NOT in state 3 at time t+1. Give interpretation of your answer. [3]
 - vii) Find the probability of path: New → Active → Dormant → Lost [2]
- [24 Marks]**
