D/E/TZ

Reg. No: 0439

VIT®
Vellore Institute of Technology

Final Assessment Test – April 2026

Course: BCSE334L - Predictive Analytics

Class NBR(s): 2245

Slot: E1+TE1

Time: Three Hours

Max. Marks: 100

- KEEPING MOBILE PHONE/ANY ELECTRONIC GADGETS, EVEN IN 'OFF' POSITION IS TREATED AS EXAM MALPRACTICE
- DON'T WRITE ANYTHING ON THE QUESTION PAPER

COs	CO Statements
CO1	Understand the importance of predictive analytics and processing of data for analysis.
CO2	Describe different types of predictive models.
CO3	Apply regression and classification model on applications for decision making and evaluate the performance.
CO4	Analyze the impact of class imbalance on performance measure for model predictions and models that can mitigate the issue during training.
CO5	Define and apply time series forecasting models in a variety of business contexts.

BL – Blooms Taxonomy Level (1 – Remember, 2 – Understand, 3 – Apply, 4 – Analyse, 5 – Evaluate, 6 – Create)

Answer ALL Questions
(10 X 10 = 100 Marks)

1. A manufacturing unit systematically captures batch-wise measurements such as operating temperature, applied pressure, machine rotational speed, and raw material grade. The management has observed a significant escalation in defective output and seeks to diagnose the underlying causes and formulate corrective measures to mitigate future quality deterioration. **CO1 BL4**

   i) Which category of analytics is most appropriate for uncovering the root causes of the observed increase in defect rates?

   ii) Following the analytical findings, what decision-oriented action can be undertaken to optimize the production process and minimize subsequent defects?

2. A 650-bed tertiary care hospital operating a high-volume Emergency Department is building an Integrated Hospital Analytics Platform (IHAP) with three objectives: **CO1 BL4**

   i) Estimate next week's emergency admission volume by day and shift, using seasonal patterns, event calendars, and historical trends.

   ii) Categorize patients into meaningful groups based on diagnosis codes, treatment pathways, length of stay, and readmission behaviour — with no predefined labels.

   iii) Determine how many ICU beds to reserve daily to minimize mortality risk within staffing budget and equipment constraints.

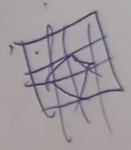
   Map each objective to its correct model type (Predictive / Descriptive / Decision). For each, justify your answer by stating the output it produces and why the other two model types cannot substitute for it.

3. A real estate dataset contains two predictors: Year Built (ranging from 1880 to 2020) and Number of Rooms (ranging from 1 to 6). CO2 BL3

i) Explain the specific problem that arises when these two predictors are used as raw values in a distance-based model like KNN.
ii) Apply standardization (Z-score) to normalize Year Built = 1990, given mean = 1960 and SD = 40. Interpret the result.
iii) Apply Min-Max normalization to Year Built = 1990 into range [0,1], given min = 1880 and max = 2020. Show working.
iv) Which normalization method would you recommend for this dataset and why?

4. a) A bank dataset contains a predictor Employment Type with four categories: Salaried / Self-Employed / Unemployed / Retired. [5] CO2 BL3

i) Create a complete dummy variable table using full one-hot encoding (k dummies) for all four categories.
ii) Now recreate the table using k−1 encoding, dropping "Unemployed" as the reference category.
iii) Explain what the k−1 rule means for a customer who is "Unemployed" — how does the model identify them without a dedicated dummy column?



b) You are a data scientist at a telecom company. You have 5 years of hourly call-volume data. Design a complete end-to-end time series analysis pipeline: (i) exploratory analysis techniques you would apply, (ii) appropriate model class selection with justification, (iii) forecasting strategy, and (iv) evaluation metrics. Identify at least two domain-specific challenges. [5] CO5 BL4

5. A data scientist builds both a parametric and a non-parametric model for the same customer churn dataset. How would you decide which is more appropriate? CO2 BL2

i) Define the key difference between parametric and non-parametric models.
ii) Describe when a parametric model would be preferred.
iii) Describe when a non-parametric model would be preferred.
iv) Name one method to formally compare the two models.

6. A wearable health device records heart rate every second but the raw signal is very noisy due to movement artefacts. The engineer applies a smoothing filter. Explain the signal estimation task here. CO5 BL4

i) Describe signal estimation as it applies to this heart rate series.
ii) State what the observed series consists of in terms of signal and noise.

Name one technique used for signal estimation in noisy time series.

7. A logistics company benchmarks two delivery time prediction models. Model A yields MAE = 4, RMSE = 5. Model B yields MAE = 4, RMSE = 9. The company operates under a penalty framework where extreme delivery delays trigger disproportionately high costs. Critically evaluate both models using the given metrics and justify your deployment recommendation. CO3 BL5

i) Calculate MAE and interpret what the identical MAE scores reveal about the average error behaviour of both models.

ii) Calculate RMSE and explain why it penalises large errors more heavily than MAE, and analyse what the divergence in RMSE values implies about the distribution of residuals in each model.

iii) Justify your model recommendation by connecting the statistical properties of RMSE to the company's asymmetric penalty cost structure.

8. An insurance company derives three rules from a claims prediction tree using two predictors: ClaimHistory and VehicleAge. Rule 1: ClaimHistory >= 3 and VehicleAge >= 5 → Predict 8000. Rule 2: ClaimHistory >= 3 and VehicleAge < 5 → Predict 4000. Rule 3: ClaimHistory < 3 → Predict 1500. Analyse the rule structure. CO3 BL4

Rules vs Trees

i) State the number of splits in the original tree that produced these three rules.

ii) Explain how Rule 3 differs structurally from Rules 1 and 2.

iii) Describe one scenario where a sample could be covered by multiple rules after pruning.

iv) State one business advantage of presenting these rules over the tree diagram to stakeholders.

9.a) A fraud detection model predicts 20% fraud probability for a group of 500 transactions. The compliance team asks what well-calibrated means. CO3 BL2

i) Define well-calibrated probability in simple terms.

ii) State approximately how many of the 500 transactions should be fraudulent if well-calibrated.

iii) Name one real-world domain where well-calibrated probabilities are essential.

OR

9.b) A hospital screens 2000 patients for a rare genetic condition. Only 40 have the condition. Compute and interpret the prevalence. CO3 BL2

i) Compute the prevalence using the correct formula.

ii) Interpret what this value implies about the rarity of the condition.

iii) State one practical consequence of low prevalence for the testing programme.

10.a) A cancer detection model uses a default threshold of 0.5. A radiologist argues that missing a true cancer case is far more dangerous than a false alarm. How should the threshold be adjusted? CO4 BL5

i) State the effect of lowering the decision threshold on sensitivity.

ii) State the trade-off that occurs when the threshold is lowered.

iii) Recommend a specific direction of threshold change and justify it clinically.

OR

10.b) Three cutoff selection strategies are described: CO4 BL2

i) user-specified sensitivity/specificity target,

ii) minimum distance to perfect model,

iii) Youden's J index.

Critically compare these three approaches across the following criteria: mathematical interpretability, sensitivity to class imbalance ratio, robustness to sample size, and applicability in asymmetric cost environments.

⇔⇔⇔ D/E/TZ ⇔⇔⇔