Roll No: | | | | | | | | | | | |

PRESIDENCY UNIVERSITY

GAIN MORE KNOWLEDGE
REACH GREATER HEIGHTS

Presidency University Act, 2013 of the Karnataka Act No. 41 of 2013 | Established under Section 2(f) of UGC Act, 1956
Approved by AICTE, New Delhi | Approved By BCI
Bengaluru

Mid Term Examinations - October 2026

Date: 05/10/2026 **Time:** 11:45 AM - 01:15 PM

Course Code: CBD3407 **Course Name:** Feature Engineering And Model Selection

Semester: Seventh Semester **Max. Marks:** 50 **Weightage:** 50%

CO - Levels	CO1	CO2
Marks	46	44

PART A: Answer Following Questions. 10 M

Q.N.	Questions	M	CO	BT
1	Feature engineering is the process of **creating, transforming, selecting, or modifying features** from raw data so that they can be effectively used by a machine-learning model. One role of feature: • It helps convert raw data into **informative numerical representations** that improve the model's ability to learn patterns.	2	CO1	BT2
2	Data profiling is the process of **examining and summarizing the structure, content, quality, and characteristics of a dataset**. Any two data quality issues: • Missing values • Duplicate records • Outliers • Incorrect data types • Inconsistent values • Invalid values	2	CO1	BT2
3	**Standardization** Converts data to approximately mean 0 and standard deviation 1. Formula: z=x−μσz=\frac{x-\mu}{\sigma} **Min-Max Normalization** Rescales data to a specified range, commonly [0,1]. Formula: x′=x−xminxmax−xminx'=\frac{x-x_{min}}{x_{max}-x_{min}}	2	CO1	BT2
4	Feature extraction is the process of **deriving useful and meaningful features from raw data**. It is important because it: • Converts complex/raw data into a numerical representation. • Reduces irrelevant information. • Helps machine-learning algorithms identify useful patterns.	2	CO2	BT1
5	TF-IDF stands for **Term Frequency–Inverse Document Frequency**. It assigns a numerical weight to a word based on its importance in a document relative to a collection of documents. The two components are: 1. **TF – Term Frequency** 2. **IDF – Inverse Document Frequency** TF-IDF=TF×IDF	2	CO2	BT2

PART B : Answer Any 1 Following Questions. 10 M

Q.N.	Questions	M	CO	BT

ANSWER KEY

PTO (Please Turn Over)

6	A typical data cleaning and preprocessing workflow is: **Raw Data → Data Profiling → Identify Quality Issues → Clean Data → Handle Missing Values → Handle Outliers → Transform Data → Scale/Encode → Model-Ready Data** **1. Data Quality Issues — 2 marks** Common issues include: • Missing values • Duplicate records • Incorrect data types • Invalid values • Inconsistent formats • Outliers • Noisy data **2. Data Profiling — 1 mark** Data profiling examines: • Number of records and attributes • Data types • Missing-value percentage • Unique values • Statistical summaries • Distribution of variables **3. Data Cleaning — 2 marks** Examples: • Remove duplicate records. • Correct invalid values. • Standardize inconsistent formats. • Correct data types. • Remove irrelevant attributes where appropriate. **4. Missing Value Treatment — 1.5 marks** Possible methods: • Mean/median imputation for numerical variables • Mode imputation for categorical variables • Forward/backward filling for suitable time-series data • Model-based imputation • Deletion when appropriate **5. Outlier Treatment — 1 mark** Methods include: • IQR method • Z-score • Isolation Forest • Capping/winsorization • Transformation • Removal when justified **6. Transformation and Scaling — 1.5 marks** • Log transformation can reduce right skewness. • Power transformations can make distributions more suitable. • Standardization converts values using mean and standard deviation. • Min-Max scaling maps values to a specified range.	10	CO1	BT3

ANSWER KEY

7	**1. Meaning of Missing Data — 1 mark** Missing data occurs when the value of an attribute is unavailable or not recorded for an observation. **2. Numerical Imputation — 3 marks** **Mean:** xmissing=∑xinx_{missing}=\frac{\sum x_i}{n} Suitable when the data is reasonably symmetric and does not contain strong outliers. **Median:** Replace missing values with the median. Useful for skewed data or data containing outliers. **Model-based imputation:** Predict missing values using other features. Examples: • Regression • KNN • Machine-learning models **3. Categorical Imputation — 2 marks** **Mode imputation:** Replace missing values with the most frequent category. Example: **Gender** Male Female Male Missing Mode = Male → Missing value becomes Male. Another option is assigning an explicit category such as **"Unknown"** when missingness itself is meaningful. **4. Time-Series Data — 1 mark** For suitable ordered data: • Forward fill • Backward fill • Interpolation **5. Deletion — 1 mark** Rows or columns may be removed when the amount of missing data is very high and deletion does not introduce significant bias. **6. Selection of Technique — 2 marks** **Attribute/Data Situation — Suitable Method** Symmetric numerical data — Mean Skewed numerical data — Median Categorical data — Mode Missingness has meaning — "Unknown" category Time-series data — Forward fill/interpolation Complex relationships — Model-based imputation	10	CO1	BT3

ANSWER KEY

PART C: Answer Any 1 Following Questions. 10 M				
Q.N.	**Questions**	**M**	**CO**	**BT**

8	**1. Definition — 1 mark** An outlier is an observation that is **significantly different from the general pattern of the data**. **2. Statistical Methods — 2 marks** Common methods include: **Z-score:** $Z=x-\mu\sigma Z=\frac{x-\mu}{\sigma}$ A value with a sufficiently large absolute Z-score may be considered an outlier. **IQR method:** $IQR=Q3-Q1IQR=Q_3-Q_1$ Lower boundary: $Q1-1.5(IQR)Q_1-1.5(IQR)$ Upper boundary: $Q3+1.5(IQR)Q_3+1.5(IQR)$ **3. Numerical Example — 4 marks** Consider: $10,12,13,14,15,16,18,5010,12,13,14,15,16,18,50$ Suppose: $Q1=12,Q3=18Q_1=12,\quad Q_3=18$ Therefore: $IQR=18-12=6IQR=18-12=6$ Lower boundary: $12-(1.5\times6)=312-(1.5\times6)=3$ Upper boundary: $18+(1.5\times6)=2718+(1.5\times6)=27$ Any value below 3 or above 27 is an outlier. Therefore: $50>2750>27$ So **50 is identified as an outlier**. **4. Handling Outliers — 3 marks** Possible strategies: • Remove the observation if it is an error. • Cap/winsorize the value. • Apply a transformation such as log transformation. • Use robust statistical methods. • Retain it if it represents a genuine and important observation.	10	CO1	BT3

ANSWER KEY

9	**1. Need for Transformation — 2 marks** Data transformation changes the scale or distribution of a feature. It can: • Reduce skewness. • Reduce the influence of extreme values. • Make patterns easier for some models to learn. • Improve numerical stability. **2. Log Transformation — 3 marks** A common transformation is: x′=log(x)x'=\log(x) For data containing zero, a commonly used version is: x′=log(1+x)x'=\log(1+x) Example: **Original log(1+x) approximately** 0 0 9 2.30 99 4.60 999 6.91 Large values are compressed more strongly than small values. **3. Power Transformation — 3 marks** Power transformations modify the feature using a power parameter. General form: x′=xλx'=x^\lambda Examples: • λ=0.5\lambda=0.5 → square-root transformation • λ=1\lambda=1 → no transformation • λ=−1\lambda=-1 → reciprocal transformation Methods such as **Box-Cox** and **Yeo-Johnson** can determine an appropriate transformation parameter. **4. Application — 2 marks** For a highly right-skewed income feature, a logarithmic transformation can compress extremely high income values and make the distribution less skewed.	10	CO1	BT3

PART D: Answer Any 1 Following Questions. 10 M				
Q.N.	**Questions**	**M**	**CO**	**BT**

ANSWER KEY

10	**1. Numerical Feature Construction — 1 mark** It is the process of creating new numerical features from existing variables to represent useful information. **2. Aggregation — 3 marks** Aggregation combines multiple observations into a summary value. Example: A customer has: **Transaction Amount** T1 500 T2 800 T3 700 Features can be created: TotalAmount=500+800+700=2000TotalAmount=500+800+700=2000 AverageAmount=20003=666.67AverageAmount=\frac{2000}{3}=666.67 Other aggregate features: • Minimum • Maximum • Count • Standard deviation **3. Binning — 3 marks** Binning converts continuous numerical values into intervals. Example: **Age Age Group** 18–30 Young 31–50 Middle 51–70 Senior This can simplify continuous information and capture nonlinear patterns. **4. Mathematical Transformations — 3 marks** New features can be created using mathematical operations. Example: Given: Height=1.7mHeight=1.7m Weight=70kgWeight=70kg BMI can be constructed: BMI=WeightHeight2BMI=\frac{Weight}{Height^2} BMI=701.72=24.22BMI=\frac{70}{1.7^2}=24.22 Other transformations include: • Square • Square root • Log • Ratios • Differences	10	CO2	BT3

ANSWER KEY

11	**1. Definition — 2 marks** Categorical feature engineering converts categorical values into numerical representations suitable for machine-learning algorithms. **2. Example — 2 marks** Consider: **Customer City** A Chennai B Chennai C Bangalore D Chennai E Mumbai Category counts: • Chennai = 3 • Bangalore = 1 • Mumbai = 1 **3. Count Encoding — 3 marks** Each category is replaced by its frequency/count. **City Count Encoding** Chennai 3 Chennai 3 Bangalore 1 Chennai 3 Mumbai 1 **4. Frequency Encoding — 3 marks** Each category is replaced by its relative frequency. Frequency=Count(category)Total observationsFrequency=\frac{Count(category)}{Total\ observations} Therefore: Chennai=35=0.6Chennai=\frac{3}{5}=0.6 Bangalore=15=0.2Bangalore=\frac{1}{5}=0.2 Mumbai=15=0.2Mumbai=\frac{1}{5}=0.2 **City Frequency** Chennai 0.6 Bangalore 0.2 Mumbai 0.2	10	CO2	BT3

PART E: Answer Any 1 Following Questions. 10 M

Q.N.	Questions	M	CO	BT

ANSWER KEY

12	Consider two documents: • **D1:** "cat likes milk" • **D2:** "dog likes milk" **1. Tokenization — 2 marks** Words: D1: cat, likes, milkcat,\ likes,\ milk D2: dog, likes, milkdog,\ likes,\ milk **2. Vocabulary — 2 marks** Unique words: [cat, dog, likes, milk][cat,\ dog,\ likes,\ milk] **3. Constructing Vectors — 4 marks** Count the occurrence of each vocabulary word. **Document cat dog likes milk** D1 1 0 1 1 D2 0 1 1 1 Therefore: D1=[1,0,1,1]D1=[1,0,1,1] D2=[0,1,1,1]D2=[0,1,1,1] **4. Advantages and Limitation — 2 marks** **Advantages:** • Simple • Easy to implement • Converts text into numerical vectors **Limitation:** • Does not capture word order or semantic meaning. • Vocabulary size can become very large.	10	CO2	BT3

ANSWER KEY

13	**1. Image Feature Extraction — 2 marks** Image feature extraction is the process of converting an image into a set of **numerical descriptors** that represent important visual characteristics. **2. Color Features — 2 marks** Color can be represented using: • RGB • HSV • Grayscale A **color histogram** represents the frequency of different intensity/color values. For grayscale: H(k)=number of pixels having intensity kH(k)=\text{number of pixels having intensity }k The histogram can be normalized: Hn(k)=H(k)NH_n(k)=\frac{H(k)}{N} where NN is the total number of pixels. **3. Texture Features — 3 marks** Texture describes patterns such as: • Smoothness • Roughness • Regularity • Contrast Methods include: • GLCM • Local Binary Pattern (LBP) • Gabor filters GLCM-based descriptors include: • Contrast • Homogeneity • Energy • Correlation **4. Shape Features — 2 marks** Shape describes the geometric structure of objects. Possible descriptors: • Area • Perimeter • Aspect ratio • Circularity • Contour descriptors • Hu moments **5. Numerical Representation — 1 mark** The extracted values can be combined into a feature vector: F=[f1,f2,f3,...,fn]F=[f_1,f_2,f_3,\ldots,f_n] This vector can then be supplied to a machine-learning model.	10	CO2	BT2

ANSWER KEY