CSM-355: Machine Learning Project

CA-1 [30 Marks]

Category	Evaluation Criteria	Marks Allocation (40 Total)
Problem Understanding & Definition	- Clarity of problem statement - Justification for solving the problem - Defined objectives & hypotheses	10 marks
Dataset Selection & Preprocessing	- Dataset relevance and quality - Handling of missing values, outliers, data normalization - Feature selection/engineering	15 marks
Viva		5 marks

Instructions:

1. Report Format: Submit your project report in PDF format only. Submissions in any other format will not be accepted.
2. Screenshots & Visuals: Include relevant screenshots wherever necessary (e.g., data cleaning methods, visualizations, model performance, etc.).
3. Deadline: Late submissions will not be accepted, and you will receive a zero (0) score if you fail to submit on time.
4. Reference Sample: A sample report has already been provided. Refer to it to ensure proper formatting and content coverage.
5. Ensure you follow these guidelines strictly to avoid any issues with your submission.

*** Sample Report**

1. Problem Understanding & Definition

1.1 Clarity of Problem Statement (4 Marks)

- Clearly define the problem you are solving.
- The problem statement should be concise, specific, and explain what issue you are addressing.
- Avoid vague or overly broad descriptions. Instead, focus on a well-defined issue.
- Provide real-world context. Who faces this problem? Why is it important?

Example: Good Problem Statement:

"This project aims to build a machine learning model to predict customer churn in an e-commerce business using transactional and behavioral data. By identifying customers likely to leave, businesses can take proactive retention measures."

1.2 Justification for Solving the Problem (3 Marks)

- Explain why this problem matters and who will benefit from the solution.
- Provide real-world significance.
- Include any relevant statistics or industry insights to strengthen your justification.

Example:
"Customer churn is a major concern for e-commerce businesses, as acquiring a new customer is 5 times more expensive than retaining an existing one. A machine learning model that predicts churn can help businesses implement targeted marketing strategies to improve retention rates."

1.3 Defined Objectives & Hypotheses (3 Marks)

- Clearly outline what you aim to achieve in this project.
- State the key objectives in bullet points.
- If applicable, define one or more hypotheses that your model will test.

Example:

Objectives:

- Develop a machine learning model to predict customer churn.
- Analyze key factors that contribute to churn (e.g., purchase frequency, inactivity period).
- Evaluate different ML algorithms and select the best-performing model.

Hypothesis:
"Customers with longer inactivity periods and lower purchase frequency are more likely to churn."

2. *Dataset Selection & Preprocessing*

2.1 Dataset Relevance and Quality (3 Marks)

2.1.1 Dataset Selection

- Choose a dataset that is relevant to your problem statement.
- Clearly mention the source of the dataset (e.g., Kaggle, UCI ML Repository, real-world data collection).
- Ensure the dataset is large enough and diverse to build a reliable model.
- Provide a brief description of the dataset:
    - Number of rows and columns
    - Types of features (categorical, numerical, text, etc.)

Example:
"We are using the 'Customer Churn Dataset' from Kaggle, which contains 10,000 customer records. It includes 14 features such as customer tenure, monthly charges, and contract type, which will help predict churn."

2.2. Handling Missing Values, Outliers, and Data Normalization (3 Marks)

2.2.1. Handling Missing Values

- Identify missing values using df.isnull().sum() and explain how you handle them:
    - Remove rows/columns if missing data is high (>30%).
    - Impute values using mean, median, or mode.
    - Use predictive imputation (e.g., KNN Imputer).

Example:
"The 'TotalCharges' column had 5% missing values, which we replaced with the median value."

2.2.2. Handling Outliers

- Detect outliers using boxplots, histograms, or the IQR method.
- Decide whether to remove, transform, or cap outliers.

Example:
"Using a boxplot, we identified extreme values in the 'MonthlyCharges' column and capped them at the 99th percentile to prevent model bias."

2.2.3. Data Normalization & Standardization

- Normalize numerical data (MinMaxScaler) or standardize it (StandardScaler) when required.
- Ensure categorical features are encoded properly (One-Hot Encoding, Label Encoding).

Example:
"We applied MinMax Scaling to normalize numerical features like 'TotalCharges' and used One-Hot Encoding for categorical features like 'Contract Type'."

2.3. Feature Selection & Engineering (4 Marks)

A. Feature Selection

- Identify the most important features using:
  - Correlation Matrix
  - Recursive Feature Elimination (RFE)
  - Feature Importance (from models like Random Forest)
- Remove irrelevant or redundant features.

Example:
"The 'CustomerID' column was removed as it does not contribute to churn prediction. Also, highly correlated features were eliminated to prevent multicollinearity."

B. Feature Engineering

- Create new features that improve model performance.
- Examples of feature engineering:
  - Converting date columns into time-based features (e.g., tenure in months).
  - Creating interaction features (e.g., MonthlyCharges * Tenure).
  - Grouping categorical data into broader categories.

Example:
"We created a new feature 'Average Monthly Spend' by dividing 'TotalCharges' by 'Tenure' to help the model identify spending patterns."