Parul University

Faculty of Engineering and Technology

Parul Institute of Technology

Machine Learning (303105353)

Assignment No. - 01

---

1. A hospital plans to develop a machine learning system to predict patient readmission risk based on medical history, lab results, and hospital stay duration. The team faces issues like incomplete medical records, privacy concerns, and difficulty integrating data from different hospital branches.

   Question:

   What are the major challenges in designing this machine learning system, and how can each challenge be addressed to improve system reliability and fairness?

2. A researcher develops a spam detection model and claims that it can correctly classify emails with **95% accuracy and 95% confidence**, given sufficient training data.

   Question:

   Explain how the **PAC (Probably Approximately Correct)** learning framework applies to this claim. What does the statement imply about the model's learning feasibility and generalization?

3. A weather prediction system outputs the following probabilities for tomorrow's weather:

   - Sunny: 0.6
   - Cloudy: 0.3
   - Rainy: 0.1

   If a picnic is planned only if the chance of rain is **below 20%**,

   Question:

   What is the probability that the picnic will proceed? How can such probability-based decisions improve decision-making in ML systems?

4. An insurance company collects customer data with the following issues:

   - Some "Age" values are missing.
   - "Income" values vary widely (₹20,000 to ₹20,00,000).
   - "Gender" is stored as text ("Male," "Female," "M," "F").

Question:

Describe how you would **preprocess** this data before feeding it into a machine learning algorithm. Include steps for **handling missing values, normalization, and categorical conversion**.

5. A credit card company monitors daily transaction amounts for fraud detection. For one customer, the average daily spending is ₹3,000 with a standard deviation of ₹500. One day, the customer makes a ₹6,000 purchase.

   Question:

   Using the **Z-score method**, determine whether this transaction is an outlier. How can identifying such outliers help improve fraud detection systems?

6. A data scientist is comparing multiple models (Logistic Regression, Random Forest, and SVM) for predicting customer churn. During testing, the Random Forest performs best but shows signs of overfitting.

   Question:

   Explain how **model selection** and **hyperparameter tuning** can be applied to balance performance and generalization. What evaluation metrics or validation techniques would you use to select the final model?

Date:15/12/2025

Dr. Vinod Patidar
Asso. Prof. (CSE Dept)
PIT, Vadodara

Note: Submit the assignment on or before 20/12/2025.