National University of Computer and Emerging Sciences, CFD Campus

	Course Name:	**Data Science**	**Course Code:**	
	Program:	**Computer Science**	**Semester:**	**Fall 2025**
	Submission Date:		**Total Marks:**	**100**
	Section:	**BS(CS)-All Sectiond**	**Weight:**	
	Exam Type:	**Project II**		

Instructions/Note:

1. **Make single .ipynb file for all Questions and use this format to name it 22F-1234_Task1.pynb.**
2. **Make a Microsoft Word file and paste all of your Python code with all possible screenshots of every task output in MS word and submit .ipynb file with word file.**
3. **Please submit your file in this format 22F-1234_L1.**
4. **Do not submit your assignment after the deadline.**
5. **Do not copy code from any source otherwise you will be penalized with negative marks.**
6. **There Will be Viva for this Assignment.**

Dataset: The Telco Customer Churn dataset consists of 7,043 customer records containing demographic information (e.g., gender, partner status), account details (e.g., tenure, contract type, payment method), and subscribed services (e.g., fiber optic internet, phone lines). The target variable is Churn, a binary categorical label ("Yes" or "No") indicating whether a customer has cancelled their service within the last month. Your primary objective is to use the 19 independent features—ranging from numerical values like MonthlyCharges to categorical ones like TechSupport—to model the probability of customer attrition, thereby identifying which clients are at risk of leaving the company. The dataset is attached with this document.

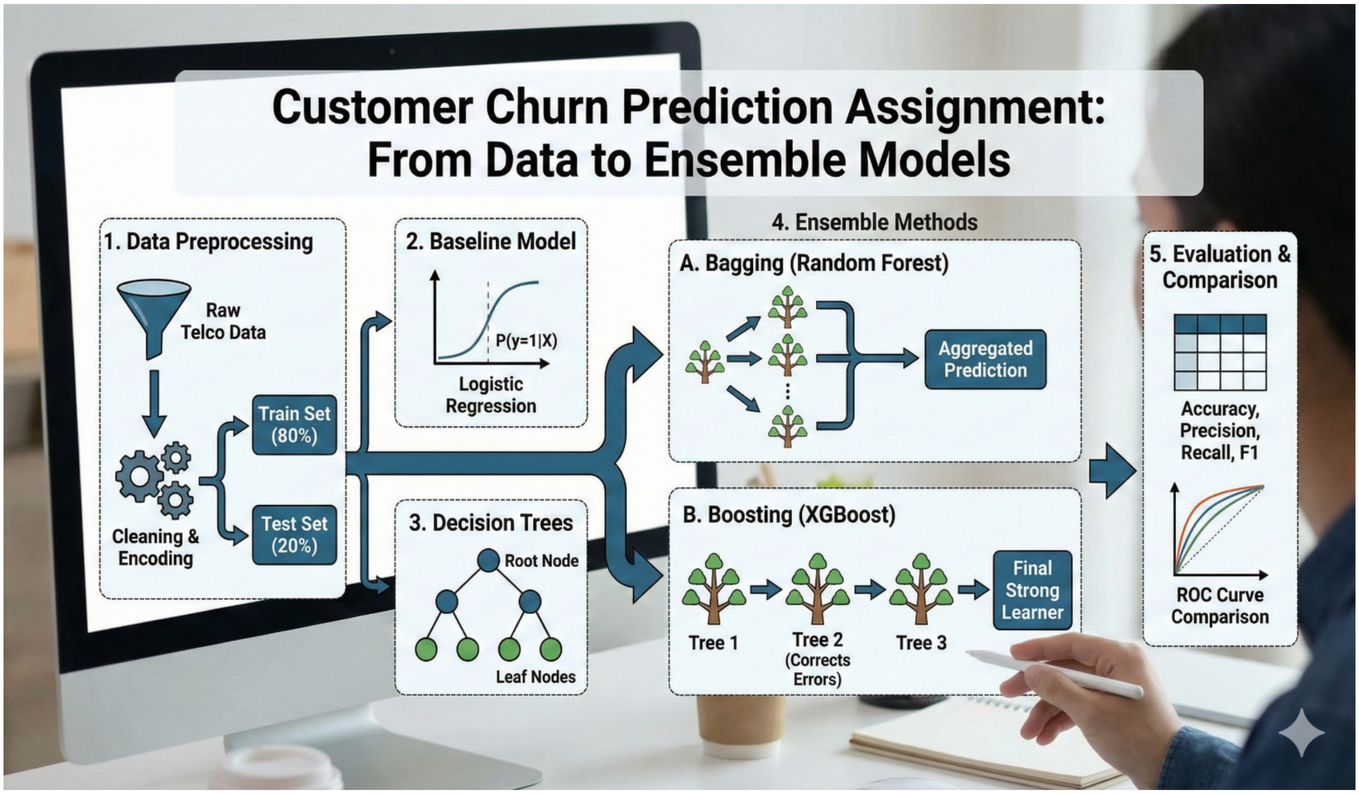
Part 1: Data Preprocessing and Feature Engineering Begin your analysis by loading the dataset and performing essential data cleaning. You will notice that the TotalCharges column, while conceptually numeric, is read as an object due to the presence of empty strings; coerce this column to a numeric format and handle any resulting missing values by dropping the affected rows. Once the data is clean, prepare your features for modeling. Convert the target variable, Churn, into a binary format (1 for Yes, 0 for No) and apply One-Hot Encoding to all categorical variables (such as Gender and InternetService), ensuring you drop the first category to prevent multicollinearity. Split your dataset into training and testing sets using an 80-20 ratio, ensuring the split is stratified to maintain the same proportion of churners in both sets. Finally, standardize your numerical features using a Standard Scaler, fitting the scaler only on the training data to avoid data leakage before transforming both sets.

Part 2: Baseline Modeling with Logistic Regression Establish a baseline for performance by training a Logistic Regression model on your scaled training data. Once the model is trained, analyze the underlying relationships by extracting the model coefficients. Visualize these coefficients in a horizontal bar chart, displaying the top features with the highest positive influence (increasing churn risk) and the top features with the highest negative influence (decreasing churn risk) to interpret the factors driving customer attrition.

Part 3: Decision Trees and Overfitting Analysis Explore non-linear relationships by implementing Decision Tree classifiers. Start by training an unconstrained Decision Tree using the entropy criterion and report its accuracy on both the training and testing sets. Next, train a second, "pruned" Decision Tree with constraints on the maximum depth and leaf size (e.g., a max depth of 5) to control model complexity. Compare the training and testing accuracies of the unconstrained tree against the pruned tree to discuss which model generalizes better and avoids overfitting. To conclude this section, visualize the pruned decision tree to interpret the decision rules it has learned.

Part 4: Ensemble Learning (Bagging and Boosting) Improve upon your single decision tree by implementing ensemble methods. First, apply Bagging by training a Random Forest Classifier with a defined number of estimators (e.g., 100 trees). Second, apply Boosting by training an XGBoost Classifier (Extreme Gradient Boosting), using logarithmic loss as the evaluation metric. Fit both models on your training data and prepare them for the final evaluation phase.

Part 5: Model Evaluation and Comparison Conclude the assignment by generating predictions on the test set for all four models: Logistic Regression, Pruned Decision Tree, Random Forest, and XGBoost. Compile a comparison table that evaluates each model's performance based on Accuracy, Precision, Recall, and F1-Score. Since failing to identify a churning customer is often more costly than a false alarm, identify the model with the highest Recall and visualize its performance using a Confusion Matrix. Finally, plot the Receiver Operating Characteristic (ROC) curves for all four models on a single graph and calculate the Area Under the Curve (AUC) for each to determine which model offers the best overall discrimination capability.


Customer Churn Prediction Assignment:
From Data to Ensemble Models
1. Data Preprocessing
Raw Telco Data
Cleaning & Encoding
Train Set (80%)
Test Set (20%)
2. Baseline Model
P(y=1|X)
Logistic Regression
3. Decision Trees
Root Node
Leaf Nodes
4. Ensemble Methods
A. Bagging (Random Forest)
Aggregated Prediction
B. Boosting (XGBoost)
Tree 1
Tree 2 (Corrects Errors)
Tree 3
Final Strong Learner
5. Evaluation & Comparison
Accuracy, Precision, Recall, F1
ROC Curve Comparison