# **Assignment: The Streamify Churn Prediction Challenge**

**Course:** AML

**Topic:** Supervised Learning (Classification)

**Due Date:** August 10th 2025

**Estimated Time:** 5 hours

### **1. The Business Problem: Retaining Customers at Streamify**

"Streamify" is a popular new video streaming service. While they are gaining new users, they are also concerned about the number of customers who cancel their subscriptions (an event known as "churn"). It is much more expensive to acquire a new customer than to retain an existing one.

Streamify's management wants to be proactive. They want to identify customers who are at a high risk of churning *before* they actually cancel their subscription. This would allow them to offer targeted promotions, special content, or customer support to encourage these at-risk users to stay.

They have hired you as a machine learning consultant to analyze their customer data and build a predictive model.

**Your mission is to build a machine learning model that can accurately predict whether a customer will churn next month.**

### **2. The Core Objective**

You will perform a full machine learning workflow:

1. **Generate the Dataset:** Run the provided Jupyter Notebook to create your unique dataset.
2. **Explore and Clean:** Investigate the dataset to understand its characteristics and handle any issues.
3. **Preprocess:** Transform the data into a format suitable for machine learning algorithms.
4. **Build and Train:** Develop at least two different classification models to predict customer churn.
5. **Evaluate:** Use appropriate metrics to measure the performance of your models and select the best one.
6. **Report:** Summarize your findings and provide actionable recommendations to Streamify's management team.

### **3. Project Tasks**

Complete the following tasks and document your entire process in a single Jupyter Notebook.

**Task 1: Data Generation (1 Mark)**

* You are be provided with a Jupyter Notebook file named churn\_data\_generation.ipynb.
* **Open this file and run all the cells.** This script will generate a file named streamify\_churn.csv.
* This streamify\_churn.csv file is the dataset you will use for all subsequent tasks. This step ensures everyone is working with the same data structure but allows for a self-contained project setup.

**Task 2: Data Exploration and Visualization (7 Marks)**

* In a *new* Jupyter Notebook, load the streamify\_churn.csv file you just created.
* Perform an initial analysis (.info(), .describe(), .isnull().sum()) to understand the data's structure.
* Visualize the data to find insights:
  + Create a bar chart to see the distribution of the target variable (will\_churn). Is the dataset balanced?
  + Create histograms for numerical features (age, monthly\_watch\_hours).
  + Create bar charts for categorical features (subscription\_plan).
  + Use a correlation heatmap to see how numerical features relate to each other.

**Task 3: Data Preprocessing (5 Marks)**

* **Handle Categorical Data:** Convert categorical features like subscription\_plan into a numerical format using One-Hot Encoding.
* **Feature Scaling:** Apply a scaler (like StandardScaler or MinMaxScaler) to your numerical features. Explain why this step is important.
* **Data Split:** Divide your data into a training set and a testing set.

**Task 4: Model Building and Training (5 Marks)**

* **Model 1 (Baseline):** Train a **Logistic Regression** model.
* **Model 2 (Advanced):** Train a more complex model, such as a **Random Forest Classifier** or something else you would want to try.
* Train the models on your training data.

**Task 5: Model Evaluation (5 Marks)**

* Make predictions on your test data using both trained models.
* Calculate the **Accuracy, Precision, Recall, and F1-Score** for each model.
* **Justify Your Choice:** In the context of this problem, which metric is most important? Is it more costly for Streamify to mistakenly predict a customer will stay (false negative) or to mistakenly predict a customer will leave (false positive)? Based on your reasoning, select the best model.

**Task 6: Final Report and Recommendations (2 Marks)**

* **Feature Importance:** If you used a Decision Tree or Random Forest, plot the feature importances. What are the top 3 factors that influence customer churn?
* **Conclusion:** Summarize your project. Which model did you choose and why? What was its performance?
* **Actionable Advice:** Provide 2-3 clear, data-driven recommendations for Streamify's management.

### **4. What to Submit**

You will submit **one** file:

1. Your main analysis **Jupyter Notebook (.ipynb)** file that includes all your code for tasks 2-6, cell outputs, and markdown explanations.