Code Assignment **CAP-1**

**Capstone Assignment 1**

**20 marks**

You could reuse much of the code that you made for the previous assignments.

**Background**

You were recently assigned the role of a data analyst in your company. Your company is a customer facing company and the number of subscribers to the company services decides the profit of the company. Your manager recently heard that you have taken a course in machine learning from Amrita. She asks you to give her a report on how to use machine learning in identifying if a customer is going to end their subscription or not. So, you decide to prepare a report to submit to your manager so that she can easily understand the abilities of ML in solving this task. Reading this report, your manager will decide whether or not she wants to use ML in for her company as well. She will also decide whether or not you are a good fit to execute this project in the company.

**Introduction**

In this assignment, you will apply the machine learning concepts that you have learned in your entire course to a real-world problem. You can use a variety of machine learning algorithms, including linear and logistic regression, SVM, ANN, and k-means clustering, to build models that can predict the outcome of a given event.

**Problem Statement**

The problem that you will be working on is to predict whether a customer will churn (cancel their subscription / account) based on their past behavior. You will need to use a dataset that contains information about customers - for example their age, gender, location, and purchase history or similar features. You will use this data to build models that can predict whether a customer is likely to churn (cancel their subscription / account).

**Model Evaluation Criteria**

Your model will be evaluated on its accuracy, precision, and recall. Accuracy is the percentage of predictions that were correct. Precision is the percentage of positive predictions that were actually positive. Recall is the percentage of positive instances that were correctly predicted as positive.

**Instructions**

1. Import the data into a Python environment.
   1. You may select any of the customer churn datasets available online. You are recommended to use any ONE of the following:
      1. <https://www.kaggle.com/datasets/blastchar/telco-customer-churn>
      2. <https://www.kaggle.com/datasets/gauravtopre/bank-customer-churn-dataset>
      3. <https://data.world/hotwan2/customer-churn-rate>
2. Split the data into a training set and a test set.
3. Choose a machine learning algorithm and train the model on the training set.
4. Evaluate the model on the test set.
5. Repeat steps 3 and 4 with different machine learning algorithms.
6. Choose the best model and explain why you chose it.
7. Use the model to make predictions on new data. (optional).
8. Write the report.

**What to Submit ?**

Upload a report in .docx format that includes all the details needed for your manager to understand what you have done, and evaluate your abilities in ML, as well as understand the capabilities of ML in solving this task. The report should finally enable your manager to decide whether you can be assigned the task of using ML for customer churn prediction model building using your company data.

You also need to upload the .**ipynb Notebook file.**

**Grading Scheme (Total 20 Marks)**

Your assignment will be graded on the following criteria:

1. **Report**: 5 Marks (note: Manager prefers succinct, to the point information, presented in a clear language, devoid of large paragraphs. She gives a lot of importance to attention to detail, and graphs / plots / tables need to be self explanatory or with inferences.)  
     
   *Note: 5 marks for extremely well structured, meaningful, and easily understandable doc, and lesser marks for progressively lower quality reports.*
2. **Graphs, plots, tables :** 3Marks
3. **Reporting of accuracy / precision / recall of the model:** 3 Marks
4. **Inferences and conclusions :** 3 Marks
5. **Code:** 5 Marks
6. **The X factor:** 1 Mark

**Time - 7-8 hours max.***(Guideline: 1 hr to understand the assignment, select and know about the dataset, and plan your assignment. 1.5 hrs per model. 1.5 hrs for report preparation). You will be using 5 hours of theory + lab time (Wed+Thursday) and have to spend 2 hours extra apart from that to complete this assignment.*

**Submission on Teams by 11.30 PM, Saturday 10 June 2023.**

**WP students: by 11.30 PM, Sunday 11 June 2023.**Late submissions are allowed till 13 June 2023 (11.30 PM) although it will result in a reduction of 20% marks.

**Plagiarism Policy**

**You are allowed to discuss.** However, you cannot copy any part of your report from another person or from the web. You cannot copy paste text/image from the internet or other sources. All reports are going to be put through TurnitIn, and any similarity score of more than 40% will be flagged as plagiarism. The penalties are as follows.

* 1st instance of plagiarism - **0 marks for this assignment**
* 2nd instance of plagiarism - **GRADE DOWN for the entire course**
* 3rd instance of plagiarism - **FAIL in the course**

By submitting this assignment, you/your group agree to the above policies on grading, submission and plagiarism, and the person submitting on behalf of the group is assumed to have taken consent from their group members.