Objectives:

- To document project components in to final report

Instructions:

- This is a group project (Team information already in D2L)
- Submit a single PDF file. Your document should be named 'Final Report' with student names on the first page
- One submission per group

For this submission, you will put all your work from week-14 to week-16 together into your final project report. This is a group assignment and only one submission per group is required. The tasks that you need to perform to complete this assignment are as following:

1. List the names of team members and the contributions of each team member.
2. Your report should contain the following sections:
    a. Introduction (i.e., description about your project, problem statement, motivation, and proposed solution).
    b. Background (i.e., literature search).
    c. Experiment Design (i.e., information about RQs, data sets, algorithms etc., that you worked in week-14)
    d. Experiment Procedure (refer to your work in week-15)
    e. Results and Discussions (Refer to a sample description below, i.e., within this document, to know how to document this section).
    f. Conclusions
    g. References
3. Submit this assignment in D2L before the due date.

You can refer to the information below on how to write the experiment design, procedure, results, and discussion for your project. The information below is just for your reference.

WRITING EXPERIMENT DESIGN

1. Identify at least two (2) Research Questions (RQs).
   [**HINT**: Research questions describes the research tasks that you will investigate in this project. They aim to run an experiment to prove/disprove your hypothesis. For example, if your group project topic is *sentiment analysis of textual reviews* then one possible RQ can be "**What are the best machine learning (ML) algorithms to calculate the**

sentiment polarity of textual reviews?", another example of a RQ can be "**What factors affect the automated generation of sentiment polarity?**"]

2. Identify the dataset description and your rationale to select it for your project. For the sentiment polarity (i.e., negative or positive) project example above, your data description will consist of something like the following:
    a. The name and location of the dataset. E.g., "*the dataset that we are using for this experiment is Amazon customer review dataset generated from year 2017- 2019 and is available at "www.amazon.com/abc/bcd. The data set consisted of 10000 reviews. This data set is selected because of large number of publicly available reviews. The reviews are also given by distinctive users and poses very less to little bias.*" You can choose any relevant details to put in this section for your data description.
    b. Note: The above description is provided as an example to you and may vary significantly for your project. This description is provided to you to give you an idea on what to write in this section.
3. Describe preprocessing of your dataset. Why was it necessary and how it helped with your RQs.
4. Describe Hardware/software descriptions. E.g., the configuration of the H/W on which this experiment is planned to run and the software's that are used. This step is important to provide the details for anyone who later wish to replicate your experiment.
5. Describe evaluation metrics and the procedure. Here, you need to identify how will you evaluate the success of your experiment or evaluate your results. E.g., the analysis will be performed using accuracy, precision or recall evaluation metrics. The comparison of success of results (e.g., automated sentiment polarity) is going to be evaluated against the sentiment polarity generated manually by domain experts.

WRITING EXPERIMENT PROCEDURE

1. For the research questions that you identified in previous assignment, describe your experiment procedure and evaluation plan. Experiment procedure describes step by step details of your approach.
2. For the experiment procedure in step-2, write the pseudocode/algorithm of your approach.

WRITING RESULTS AND DISCUSSIONS

For this part of your group project, you will elaborate on various components of your projects (as listed below). The tasks that you need to document results and discussions section is as following:

1. For the research questions (RQ's) that you identified for your project, you need to describe your results. For your understanding about this task, a hypothetical example is considered that attempts to *automate the loan eligibility of a person* depending upon various constraints passed to the prediction models. These constraints are age, income, student status, other monthly expenses, zip code, and employment status, etc. Assume that for this example, the following were identified as part of your experiment design.
    a. **Research Question (RQ-1)**: What is the accuracy of our model at predicting the eligibility of a person for a loan?
    b. **Data Set**: Assume that the data set (e.g., XYZ) used for this problem consists of real world data and there are 30,000 data instances. The data also included roughly equal number of cases that were either 'approved' or 'denied'. The models is trained using 20,000 instances and the remaining 10,000 are used for testing the prediction.
    c. **Algorithm**: Assume your model used *Probability based algorithms* to predict the outcome.
    d. **Evaluation metrics**: Your model is going to be evaluated using accuracy, recall, and precision and compared against the results from using the regression models by various researches.

    For the above experiment design (*note that this experiment design is hypothetical and contains very little description about various elements*). You need to write the result section as following:

Results

Description: The results are obtained using the probabilistic algorithm using the XYZ dataset. The results are evaluated using the RQ's identified earlier and described in XXX section (*you can refer to your previous assignment or the section in this assignment if re-writing those again*). The results are presented in Table 2, where the confusion matrix (refer Table 1) is used to identify the true-positives, true-negatives, false-positives, and false-negatives. This confusion matrix is used to calculate the results in terms of accuracy, precision, and recall. (**You can add more similar details depending upon experiment design of your project).** The results obtained are structured around the RQ's identified earlier and are as following:

Table 1. Confusion Matrix

	Actual (Yes)	Actual (No)
Predicted (Yes)	True Positive	False Positive
Predicted (No)	False Negative	True Negative

RQ-1: What is the accuracy of our model at predicting the eligibility of a person for a loan?

Using the confusion matrix (shown in Table 2), following are a few observations regarding the results. (**NOTE: Basically in this section, you will be observing the results data**)

a. The number of truly predicted instances are 8,500 out of 10,000 making the accuracy of our model to 85%.
b. There were a total of 1500 data instances that were incorrectly identified as false-negatives and false-positives.
c. It is also observed from the data that there are 5100 instances that are 'eligible' and 900 'ineligible' instances.

Table 2. Results

	Actual (Eligible = Yes)	Actual (Eligible = No)
Predicted (Eligible = Yes)	4,500	900
Predicted (Eligible = No)	600	4,000

2. Next, for the results section that you completed in step-2, you will write the discussion section. **This section basically provides the insights/reasoning about the results** obtained (i.e., the observations that you identified about results). The following write-up attempts to present how a discussion section can be written.

Discussion of Results

The discussion about the results and the major observations are presented in this section around the RQ's that are identified earlier for this project. The prominent insights and implications are described as following.
RQ-1: This RQ attempts to evaluate the performance of our model used for this project problem. The analysis is presented around the observations identified in results section earlier.

Observation-1 (Model accuracy of 85%) – The model predicted 85% of the instances accurately, which is a fair improvement over other researches presented in the literature which could only predict 79% correct instances (**NOTE: I just made this up**). One reason behind this improvement has been because of the use of probabilistic models to train our model over regression models used by the researches in literature. Probabilistic models are most suited for the type of data that is collected to study the underlying problem identified for this project. (**Next, you can also provide a reason as to why your model incorrectly predict the other 15% of the data instances**). Similarly, **you can present the reasons behind the observations from the results section and what are the implications** of these observations (*e.g., our model can be used as a substitution for regression based model, especially, to predict the loan eligibility*).