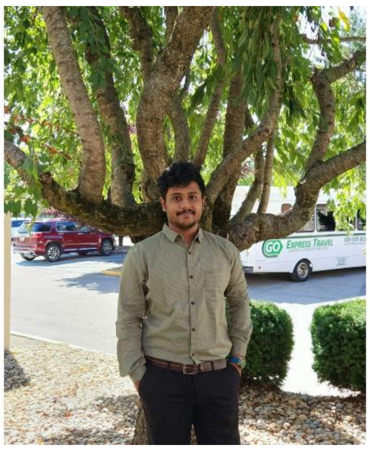
# Home Credit Default Risk (HCDR)


**Group 11**
Anuj Mahajan
Siddhant Patil
Shashwati Diware
Shubham Jambhale


## Team Members:


Shubham Jambhale
sjambhal@iu.edu


Siddhant Patil
sidpatil@iu.edu



Anuj Mahajan
anujmaha@iu.edu



Shashwati Diware
sdiware@iu.edu


## Phase Leader Plan

| Phase | Contributor | Contribution Description |
|---|---|---|
| Phase 1: Project Planning | Anuj Mahajan | Download Data, go through data, and load libraries. Create a pipeline diagram and describe the pipeline design. Describe Preprocessing, |
| Phase 1: Project Planning | Shashwati Diware | Project Abstract, ML Algorithm Names, and describe Metrics. |
| Phase 1: Project Planning | Shubham Jambhale(Phase Leader) | Understanding the problem statement, and writing table descriptions. Schedule meetings, coordinate tasks, plan phase |
| Phase 1: Project Planning | Siddhant Patil | Machine Learning Pipeline Steps and describes pipeline components. |
| Phase 2: Base Line Modelling and EDA | Anuj Mahajan (Phase Leader) | Creating Block Diagram EDA and one slide of the presentation. Schedule meetings, coordinate tasks, plan phase |
| Phase 2: Base Line Modelling and EDA | Shashwati Diware | Result Analysis EDA and one slide of the presentation. |
| Phase 2: Base Line Modelling and EDA | Shubham jambhale | Result Analysis and two slides of the presentation |
| Phase 2: Base Line Modelling and EDA | Siddhant Patil | Result Analysis and two slides of the presentation |
| Phase 3: Hyperparameter Tuning | Shashwati Diware (Phase Leader) | Testing Accuracy matrix and Schedule meetings, coordinating tasks, the planning phase |
| Phase 3: Hyperparameter Tuning | Siddhant Patil | Create and develop code for Hyperparameter tuning |
| Phase 3: Hyperparameter Tuning | Shubham Jambhale | Run and create analysis by testing the confusion / AUC matrix. Coordinate Tasks and one slide of the presentation |
| Phase 3: Hyperparameter Tuning | Anuj Mahajan | Run and analyze Lasso and ridge regression losses. Coordinate tasks and one slide of the presentation |
| Phase 4: Final Report Generation | Siddhant Patil (Phase Leader) | Plan Phase Schedule Meetings and Coordinate Tasks, analyze and go through the final results |
| Phase 4: Final Report Generation | Anuj Mahajan | Rearrange everything and go through the final documentation, list down the final recordings |
| Phase 4: Final Report Generation | Shashwati Diware | Prepare the final presentation |
| Phase 4: Final Report Generation | Shubham Jambhale | Check everything and submit the assignment before the deadline |

# Credit Assignment Plan

## Phase 1:

| Task | Task Description | Hours spent | Assigned to | Start | End |
|---|---|---|---|---|---|
| Understanding problem statement | Go through the problem statement to understand the requirements | 6 | Shubham | 11/05/22 | 11/07/22 |
| Data Exploration | Explore and analyze the data for a better understanding | 6 | Anuj | 11/07/22 | 11/09/22 |
| Project Proposal | Creating the project proposal and preparing a basic report with Abstract, ML models, and Gantt diagram | 20 | Group | 11/09/22 | 11/14/22 |

## Phase 2:

| Task | Task Description | Hours Spent | Assigned to | Start | End |
|---|---|---|---|---|---|
| Creating Block Diagram | Creating the block diagram of the basic flow of execution. | 5 | Anuj | 11/13/22 | 11/15/22 |
| Creating Pipeline Diagram | Creating the pipeline diagram of the machine learning model from analyzing the data till the result analysis | 5 | Shashwati | 11/13/22 | 11/15/22 |
| Result Analysis | Analyzing the Result | 10 | Group | 11/26/22 | 11/30/22 |
| PowerPoint Presentation | Simultaneously prepare the PowerPoint presentation and add the analyzed data into it as per need | 10 | Group | 11/20/22 | 12/03/22 |

## Phase 3:

| Task | Task Description | Hours spent | Assigned to | Start | End |
|---|---|---|---|---|---|
| Create and develop code for hyperparameter tuning | Design and develop python helper function for hyperparameter tuning | 16 | Siddhant | 11/20/22 | 11/25/22 |
| Result Analysis | Analysis of Obtained Result | 2 | Group | 12/02/22 | 12/03/22 |
| Testing Accuracy matrix | Analyzing accuracy using accuracy matrix | 2 | Shashwati | 12/03/22 | 12/04/22 |
| Testing f1 matrix | Analyzing accuracy using Confusion/AUC matrix score | 2 | Shubham | 12/03/22 | 12/04/22 |
| Lasso And Ridge Loss Functions | Analyzing the lasso and ridge loss function | 2 | Anuj | 12/03/22 | 12/04/22 |

**Phase 4:**

| Task | Task Description | Hours Spent | Assigned To | Start | End |
|---|---|---|---|---|---|
| Final Documentation | Rearrange everything and go through the final documentation, list down the final recordings | 10 | Anuj | 12/03/22 | 12/08/22 |
| Final Results | Analyze final results obtained after the final testing | 6 | Siddhant | 12/05/22 | 12/08/22 |
| Final Presentation | Prepare the final presentation | 4 | Shashwati | 12/06/22 | 12/08/22 |
| Assignment Submission | Check everything and submit the assignment before the deadline | 1 | Shubham | 12/08/22 | 12/09/22 |


## Abstract

Based on historical credit histories and repayment trends utilizing machine learning modeling, Home Credit offers unsecured lending. A user-generated credit score is calculated using criteria like the balance that the user has maintained. As part of this project, we are predicting the customer repayment status such as if the user is a defaulter or not using machine learning pipelines and models using the datasets provided by Kaggle. The data collection includes seven separate tables that aid in determining the user status, including bureau balance, credit card balance, home credit column detection, Installments payments, POS CASH balance, and previous applications. We want to offer a pipeline for logistic regression, decision trees, and random forests in Phase 1. Along with this, we will be executing L1 lasso regression and L2 ridge regression to analyze the losses during the implementation of the algorithm. In order to accurately categorize the target variables, we will apply the strategy accuracy, confusion matrix, and AUC.


## Data and Task Description

***Data source***

We are planning to use the existing datasets provided by Kaggle.
Source: [https://www.kaggle.com/c/home-credit-default-risk/data](https://www.kaggle.com/c/home-credit-default-risk/data)

***POS_CASH_balance.csv***

This dataset gives information about previous credit information such as contract status, the number of installments left to pay, DPD(days past due), etc. of the current application.

**Table 1. POS_CASH_balance.csv**

| | SK_ID_PREV | SK_ID_CURR | MONTHS_BALANCE | CNT_INSTALMENT | CNT_INSTALMENT_FUTURE | NAME_CONTRACT_STATUS | SK_DPD | SK_DPD_DEF |
|---|---|---|---|---|---|---|---|---|
| **0** | 1803195 | 182943 | -31 | 48.0 | 45.0 | Active | 0 | 0 |
| **1** | 1715348 | 367990 | -33 | 36.0 | 35.0 | Active | 0 | 0 |
| **2** | 1784872 | 397406 | -32 | 12.0 | 9.0 | Active | 0 | 0 |
| **3** | 1903291 | 269225 | -35 | 48.0 | 42.0 | Active | 0 | 0 |
| **4** | 2341044 | 334279 | -35 | 36.0 | 35.0 | Active | 0 | 0 |

***bureau.csv***

This dataset gives information about the type of credit, debt, limit, overdue, maximum overdue, annuity, remaining days for previous credit, etc.

**Table 2. Bureau.csv**

| | SK_ID_CURR | SK_ID_BUREAU | CREDIT_ACTIVE | CREDIT_CURRENCY | DAYS_CREDIT | CREDIT_DAY_OVERDUE | DAYS_CREDIT_ENDDATE |
|---|---|---|---|---|---|---|---|
| 0 | 215354 | 5714462 | Closed | currency 1 | -497 | 0 | -153.0 |
| 1 | 215354 | 5714463 | Active | currency 1 | -208 | 0 | 1075.0 |
| 2 | 215354 | 5714464 | Active | currency 1 | -203 | 0 | 528.0 |
| 3 | 215354 | 5714465 | Active | currency 1 | -203 | 0 | NaN |
| 4 | 215354 | 5714466 | Active | currency 1 | -629 | 0 | 1197.0 |

***bureau_balance.csv***

This dataset gives information about the Status of the Credit Bureau loan during the month, the Month of balance relative to the application date, Recoded ID of the Credit Bureau credit. Each row is one month of a previous credit, and a single previous credit can have multiple rows, one for each month of the credit length.

**Table 3. bureau_balance.csv**

| | SK_ID_BUREAU | MONTHS_BALANCE | STATUS |
|---|---|---|---|
| 0 | 5715448 | 0 | C |
| 1 | 5715448 | -1 | C |
| 2 | 5715448 | -2 | C |
| 3 | 5715448 | -3 | C |
| 4 | 5715448 | -4 | C |

***credit_card_balance.csv***

This dataset gives information about financial transactions aggregated values such as amount received, drawings, number of transactions of previous credit, installments, etc. Each row is one month of a credit card balance, and a single credit card can have many rows.

**Table 4. credit_card_balance.csv**

| | SK_ID_PREV | SK_ID_CURR | MONTHS_BALANCE | AMT_BALANCE | AMT_CREDIT_LIMIT_ACTUAL | AMT_DRAWINGS_ATM_CURRENT | AMT_DRAWINGS_CURRENT |
|---|---|---|---|---|---|---|---|
| 0 | 2562384 | 378907 | -6 | 56.970 | 135000 | 0.0 | 877.5 |
| 1 | 2582071 | 363914 | -1 | 63975.555 | 45000 | 2250.0 | 2250.0 |
| 2 | 1740877 | 371185 | -7 | 31815.225 | 450000 | 0.0 | 0.0 |
| 3 | 1389973 | 337855 | -4 | 236572.110 | 225000 | 2250.0 | 2250.0 |
| 4 | 1891521 | 126868 | -1 | 453919.455 | 450000 | 0.0 | 11547.0 |

***installments_payments.csv***

This dataset gives information about payments, installments supposed to be paid, and their details. There is one row for every made payment and one row for every missed payment.

**Table 5. Installments_payments.csv**

| | SK_ID_PREV | SK_ID_CURR | NUM_INSTALMENT_VERSION | NUM_INSTALMENT_NUMBER | DAYS_INSTALMENT | DAYS_ENTRY_PAYMENT | AMT_INSTALMENT |
|---|---|---|---|---|---|---|---|
| 0 | 1054186 | 161674 | 1.0 | 6 | -1180.0 | -1187.0 | 6948.360 |
| 1 | 1330831 | 151639 | 0.0 | 34 | -2156.0 | -2156.0 | 1716.525 |
| 2 | 2085231 | 193053 | 2.0 | 1 | -63.0 | -63.0 | 25425.000 |
| 3 | 2452527 | 199697 | 1.0 | 3 | -2418.0 | -2426.0 | 24350.130 |
| 4 | 2714724 | 167756 | 1.0 | 2 | -1383.0 | -1366.0 | 2165.040 |

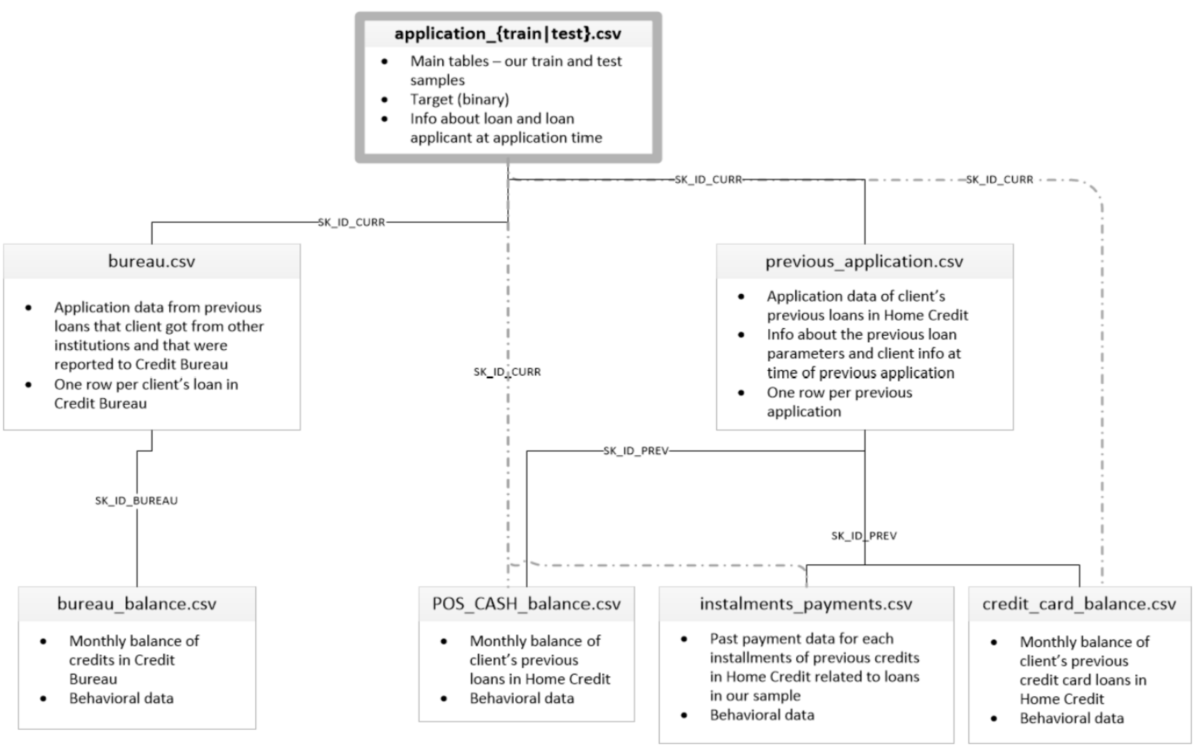
***previous_application.csv***

This dataset contains information about previous application details of an application. Each current loan in the application data can have multiple previous loans. Each previous application has one row and is identified by the feature SK_ID_PREV.

**Table 6. previous_application.csv**

| | SK_ID_PREV | SK_ID_CURR | NAME_CONTRACT_TYPE | AMT_ANNUITY | AMT_APPLICATION | AMT_CREDIT | AMT_DOWN_PAYMENT |
|---|---|---|---|---|---|---|---|
| 0 | 2030495 | 271877 | Consumer loans | 1730.430 | 17145.0 | 17145.0 | 0.0 |
| 1 | 2802425 | 108129 | Cash loans | 25188.615 | 607500.0 | 679671.0 | NaN |
| 2 | 2523466 | 122040 | Cash loans | 15060.735 | 112500.0 | 136444.5 | NaN |
| 3 | 2819243 | 176158 | Cash loans | 47041.335 | 450000.0 | 470790.0 | NaN |
| 4 | 1784265 | 202054 | Cash loans | 31924.395 | 337500.0 | 404055.0 | NaN |

**Figure 1: Data Description Diagram**



# Machine Learning Algorithm and Metrics

The outcome of this project is to predict, whether the customer will repay the loan or not. That’s why this is a classification task where the outcome is 0 or 1. To classify this problem we will be building the following machine-learning models:

1. **Logistics Regression**:
    - In our case, the number of features is relatively small i.e. <1000, and no. of examples is large. Hence logistic regression can be a good fit here for the classification.

2. **Decision Tree**:
    - Decision trees are better for categorical data and our target data is also categorical in nature that’s why decision trees are a good fit.

3. **Random Forest**:
    - Random Forest works well with a mixture of numerical and categorical features.
    - As we have a good amount of mixture of both types of features random forest can be a good fit.

# Loss Function:

- L1 Lasso Loss:
    - ➢ Less absolute shrinkage and selection operator, also known as lasso or LASSO, is a regression analysis technique that combines regularization and variable selection to improve prediction accuracy.

- L2 Ridge Loss:
    - ➢ It is a technique for making poorly stated situations regular. It is very helpful in reducing the multicollinearity issue in linear regression, which frequently arises in models with several parameters. Generally speaking, the approach improves parameter estimation problem efficiency in exchange for a manageable degree of bias.

# Metrics:

1. **Confusion Metrics:**

    - A confusion matrix, also called an error matrix, is used in the field of machine learning and more specifically in the challenge of classification. Confusion matrices show counts between expected and observed values. The result "TN" stands for True Negative and displays the number of negatively classed cases that were correctly identified. Similar to this, "TP" stands for True Positive and denotes the quantity of correctly identified positive cases. The term "FP" denotes the number of real negative cases that were mistakenly categorized as positive, while "FN" denotes the number of real positive examples that were mistakenly classed as negative. Accuracy is one of the most often used metrics in classification.

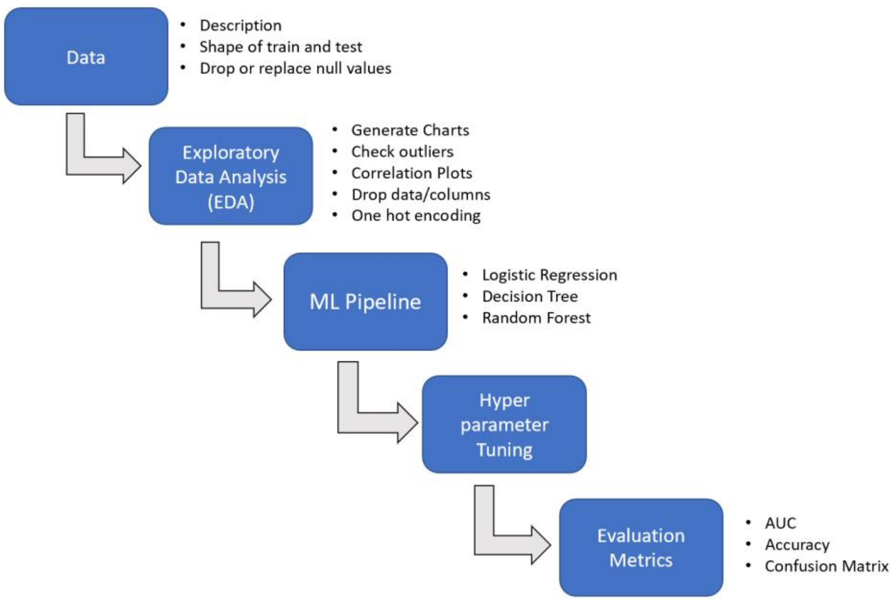
2. **AUC:**
    - AUC stands for "Area under the ROC Curve." It measures the entire two-dimensional area underneath the entire ROC curve from (0,0) to (1,1). It is a widely used accuracy method for binary classification problems

3. **Accuracy:**
   - The accuracy score is used to gauge the model's effectiveness by calculating the ratio of total true positives to total true negatives across all made predictions. Accuracy is generally used to calculate binary classification models.

     - **Accuracy Score = (True Positives + True Negatives) / (True Positives + True Negatives + False Positives + False Negatives)**

## Machine Learning Pipeline Steps:

**Figure 2: Diagram of Workflow**




### Data Preprocessing:

- Convert the raw data set into a clean data set for processing.
- First, Obtain Kaggle's raw data.
- On this Raw Data. Analyze exploratory data.

### Feature Engineering:

- Create a suitable input dataset by performing feature engineering and other processing techniques.
- Pipeline must not only select the features it wants to create from an unlimited pool of possibilities, but it must also process vast amounts of data to do so. This makes the data appropriate for the model.

**Model Selection:**

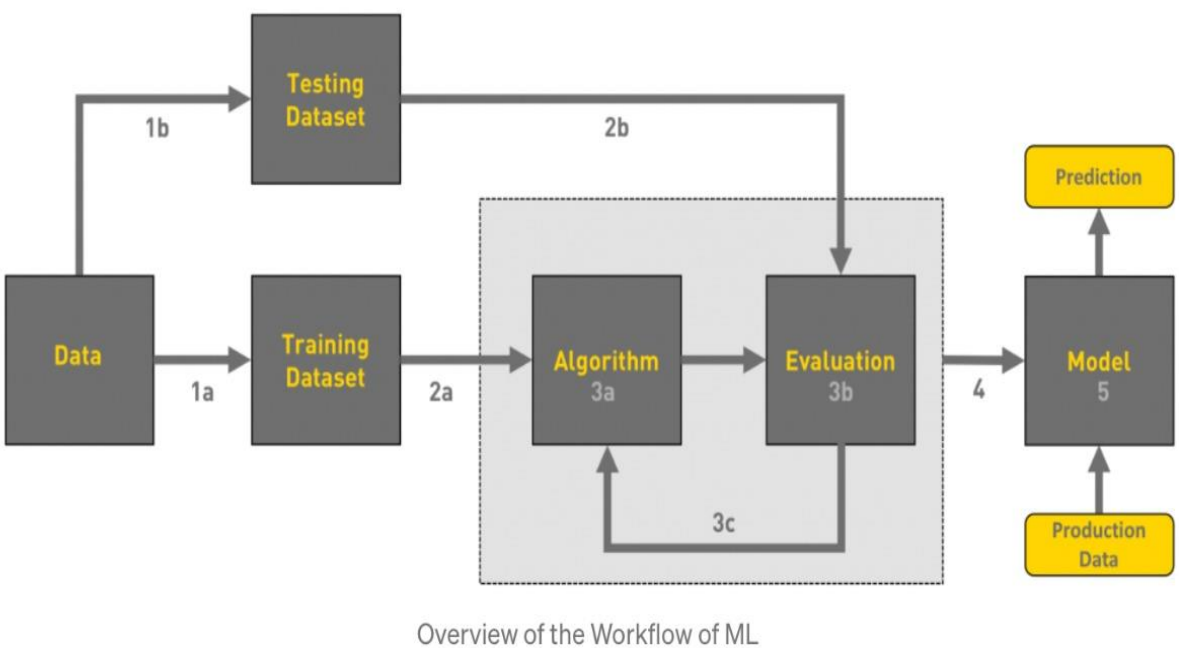
- Here, we try on different models for various option purposes.
- Develop and test several candidate models, such as Random Forest, Decision Making Trees, and Logistic Regression.
- Using the evaluation function, pick the top model with a good evaluation score.
- For this selection purposes, employ many measures for evaluation criteria, including "Accuracy," "F1 Score,".

**Prediction Generation:**

- The top performer is then chosen as the winning model when the models are tested on a new set of data that wasn't used during training.
- Once the best model has been chosen, use it to forecast outcomes based on the fresh data.
- It is then used to make predictions across all your objects.

## Block Diagram:

**Figure 3. Block Diagram**




*Referenced from: https://towardsdatascience.com/workflow-of-a-machine-learning-project-ec1dba419b94*

# Gantt Chart

**Figure 4. Gantt Chart**

## CSCI-P 556: Applied Machine Learning

FP_GroupN_ 11
Project Lead: Shubham Jambha

Project Start: Sat, 11/5/2022

Display Week: 1

| TASK | ASSIGNED TO | Phase Leader | PROGRESS | START | END |
|---|---|---|---|---|---|
| **Phase 1 Project Planning** | | Shubham | | | |
| Problem Statement Understandir | Shubham | | 100% | 11/5/22 | 11/7/22 |
| Data Exploration | Anuj | | 100% | 11/7/22 | 11/9/22 |
| Project Proposal | All | | 100% | 11/9/22 | 11/14/22 |
| **Phase 2 Baseline Modelling and EDA** | | Anuj | | | |
| Creating Block Diagram | Anuj | | 20% | 11/13/22 | 11/15/22 |
| Creating Pipeline Diagram | Shashwati | | 20% | 11/13/22 | 11/15/22 |
| Result Analysis | All | | 0% | 11/26/22 | 11/30/22 |
| Create Powerpoint Presentation | All | | 0% | 11/20/22 | 12/3/22 |
| **Phase 3 Hyperparameter Tuning** | | Shashwati | | | |
| Create and develop code for Hyp | Siddhant | | 0% | 11/20/22 | 11/25/22 |
| Result Analysis | All | | 0% | 12/2/22 | 12/3/22 |
| Testing Accuracy Matrix | Shashwati | | 0% | 12/3/22 | 12/4/22 |
| Testing Confusion/AUC Matrix | Shubham | | 0% | 12/3/22 | 12/4/22 |
| Lasso and Ridge | Anuj | | 0% | 12/3/22 | 12/4/22 |
| **Phase 4 Final Report Generation** | | Siddhant | | | |
| Final Documentaion | Anuj | | | 12/3/22 | 12/8/22 |
| Final Results | Siddhant | | | 12/5/22 | 12/8/22 |
| Final Presentation | Shashwati | | | 12/6/22 | 12/8/22 |
| Assignment Submission | Shubham | | | 12/8/22 | 12/9/22 |