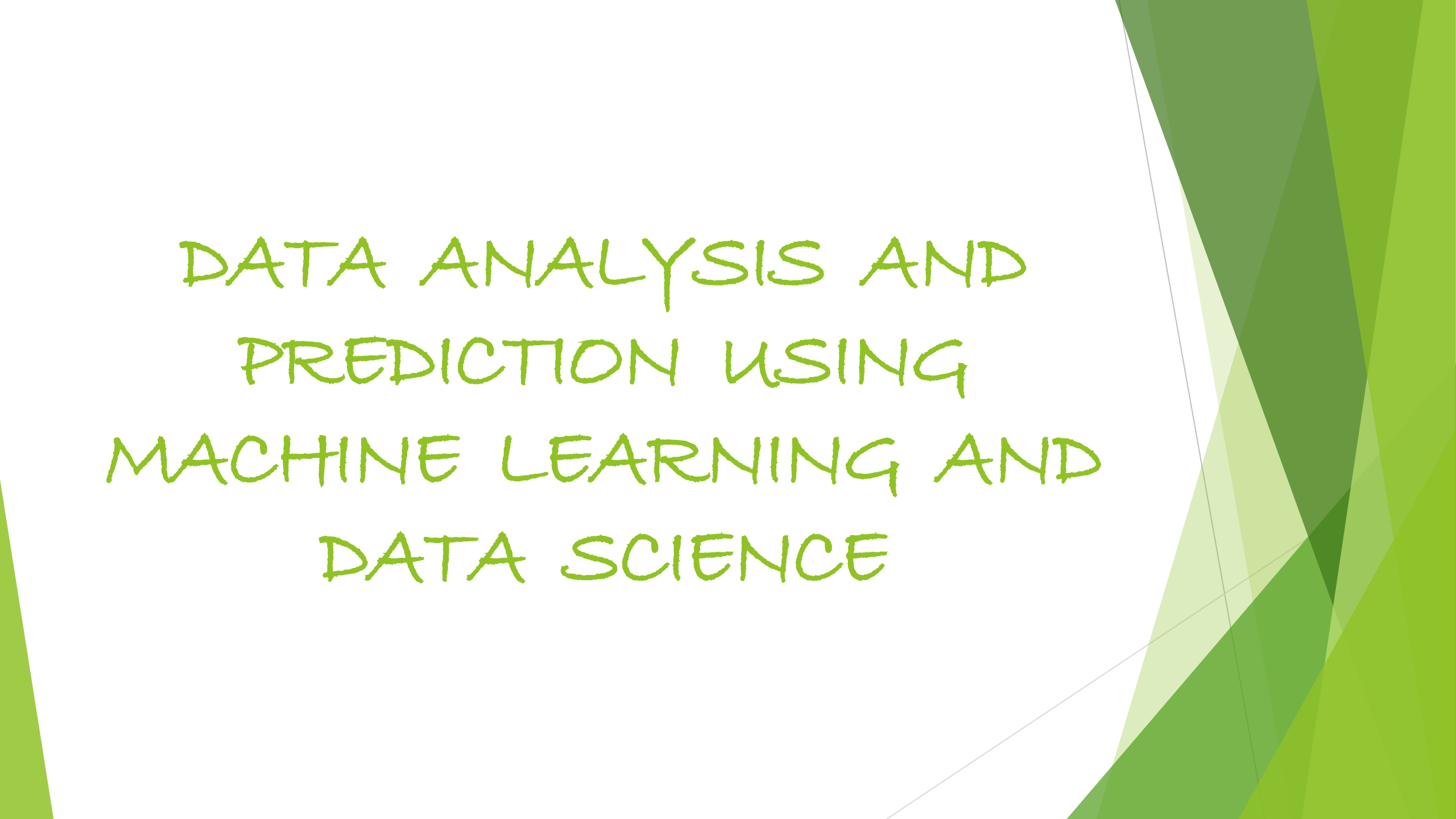




भारतीय सूचना प्रौद्योगिकी संस्थान भागलपुर
Indian Institute of Information Technology
Bhagalpur

The background features abstract, overlapping green geometric shapes, primarily triangles and polygons, in various shades of green, creating a modern and dynamic visual effect.

DATA ANALYSIS AND PREDICTION USING MACHINE LEARNING AND DATA SCIENCE

CONTENTS

- ▶ **INTRODUCTION**
- ▶ **EXPLORATORY DATA ANALYSIS**
- ▶ **MODEL DISCRIPTION & INTUTION**
- ▶ **IMPLEMENTATION STEPS**
- ▶ **CONCLUTION**
- ▶ **REQUIREMENTS OF HARDWARE & SOFTWARE**

INTRODUCTION

For this project we will be exploring publicly available data from www.lendingclub.com. Lending Club connects people who need money (borrowers) with people who have money (investors). Hopefully, as an investor you would want to invest in people who showed a profile of having a high probability of paying you back. We will try to create a model that will help predict this.

Data Attributes

- ▶ **Credit.policy:** 1 if the customer meets the credit underwriting criteria of LendingClub.com, and 0.
- ▶ **Purpose:** The purpose of the loan (takes values "credit_card", "debt_consolidation", "educational", "major_purchase", "small_business", and "all_other").

- ▶ **int.rate:** The interest rate of the loan, as a proportion (a rate of 11% would be stored as 0.11). Borrowers judged by LendingClub.com to be more risky are assigned higher interest rates.
- ▶ **installment:** The monthly installments owed by the borrower if the loan is funded.
- ▶ **log.annual.inc:** The natural log of the self-reported annual income of the borrower.
- ▶ **fico:** The FICO credit score of the borrower.
- ▶ **dti:** The debt-to-income ratio of the borrower (amount of debt divided by annual income).
- ▶ **days.with.cr.line:** The number of days the borrower has had a credit line.
- ▶ **revol.bal:** The borrower's revolving balance (amount unpaid at the end of the credit card billing cycle).
- ▶ **revol.util:** The borrower's revolving line utilization rate (the amount of the credit line used relative to total credit available).
- ▶ **inq.last.6mths:** The borrower's number of inquiries by creditors in the last 6 month.

- ▶ **delinq.2yrs:** The number of times the borrower had been 30+ days past due
- ▶ **pub.rec:** The borrower's number of derogatory public records (bankruptcy filings, tax liens, or judgments).

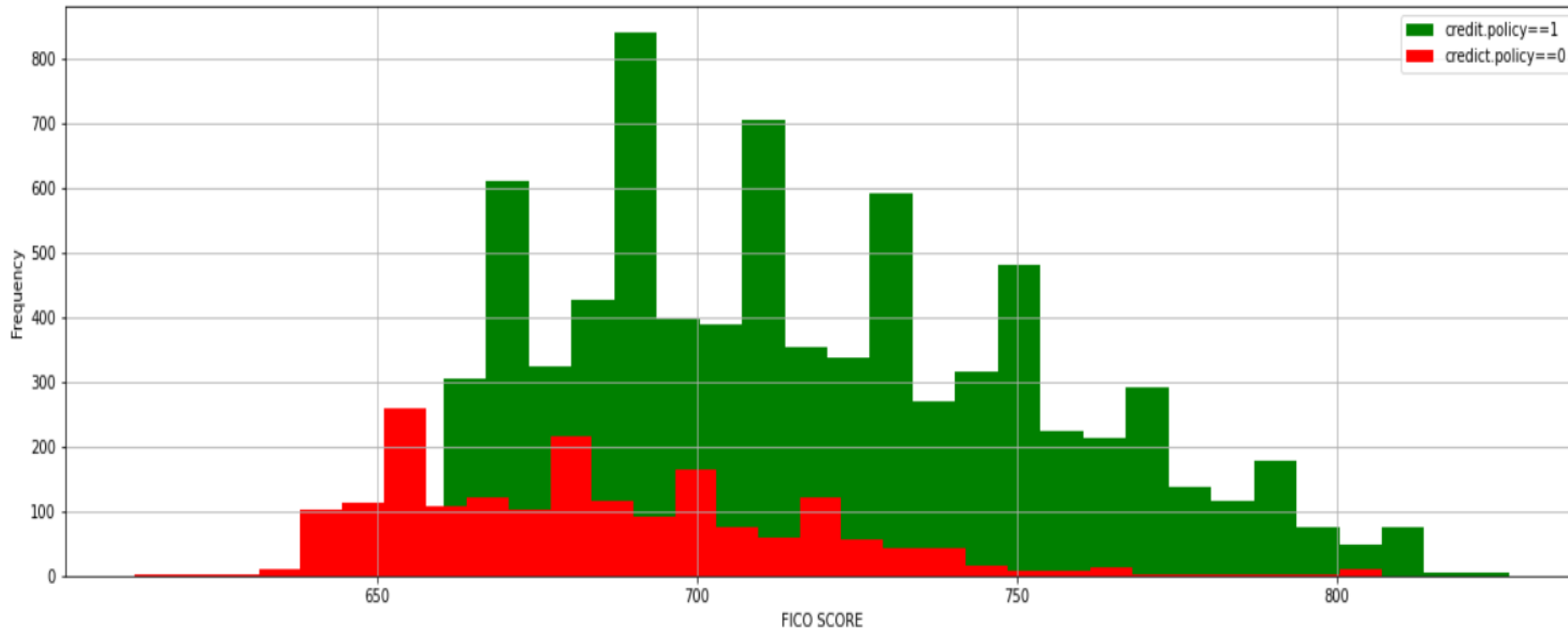
Problem Definition

Situation | Problem definition: This project is all about predicting whether a borrower will return the money to the investor or not on the basis of borrower profile which contains almost thirteen different attributes like interest rate, faco score, the purpose of taking the loan, and many more. Once we predicted that the borrower is viable. Then the company will do the disbursement of the loan.

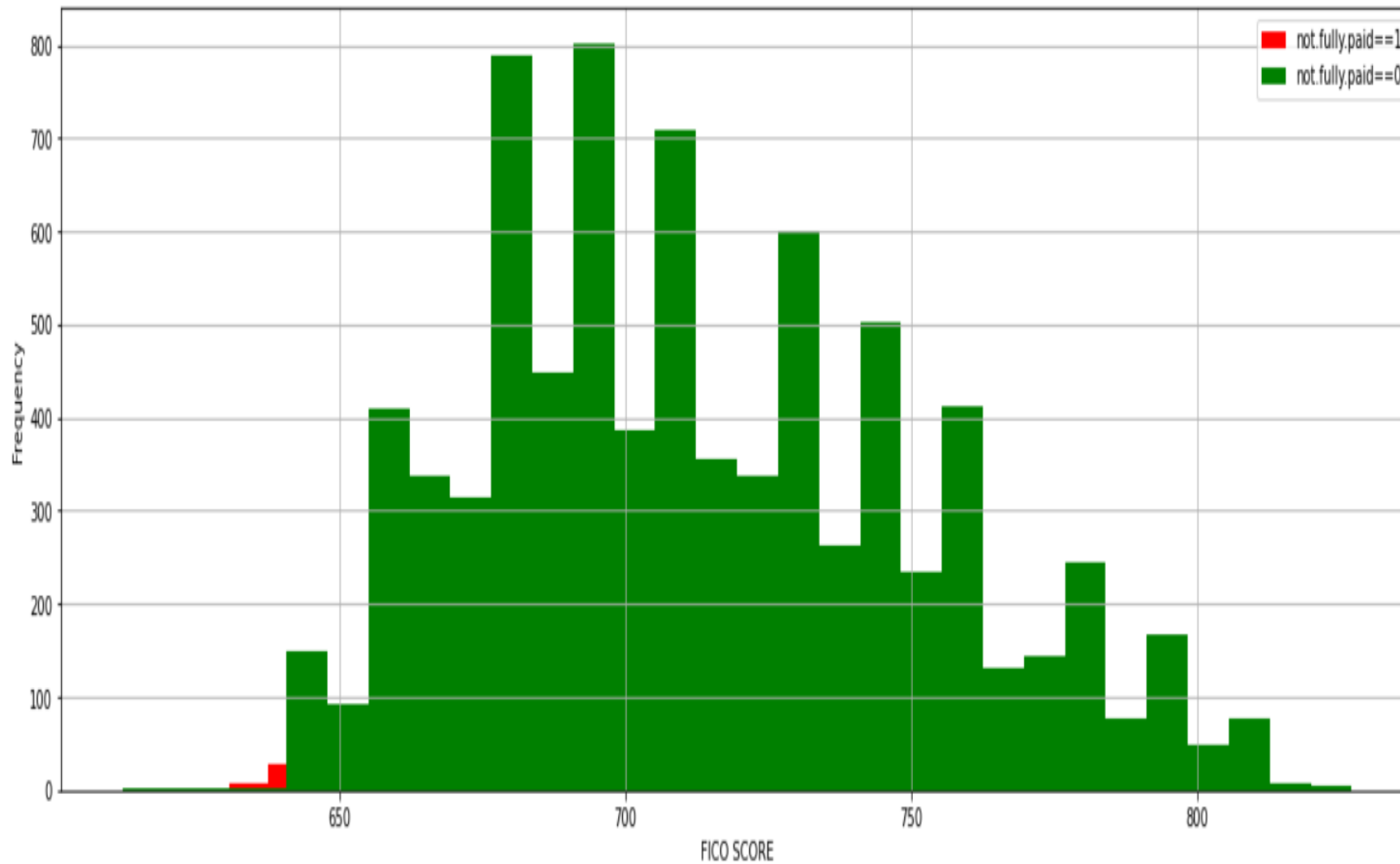
Task: Now the task was to come up with an efficient machine learning model which will result in maximum accuracy on a given data set.

Graph Exploratory Analysis

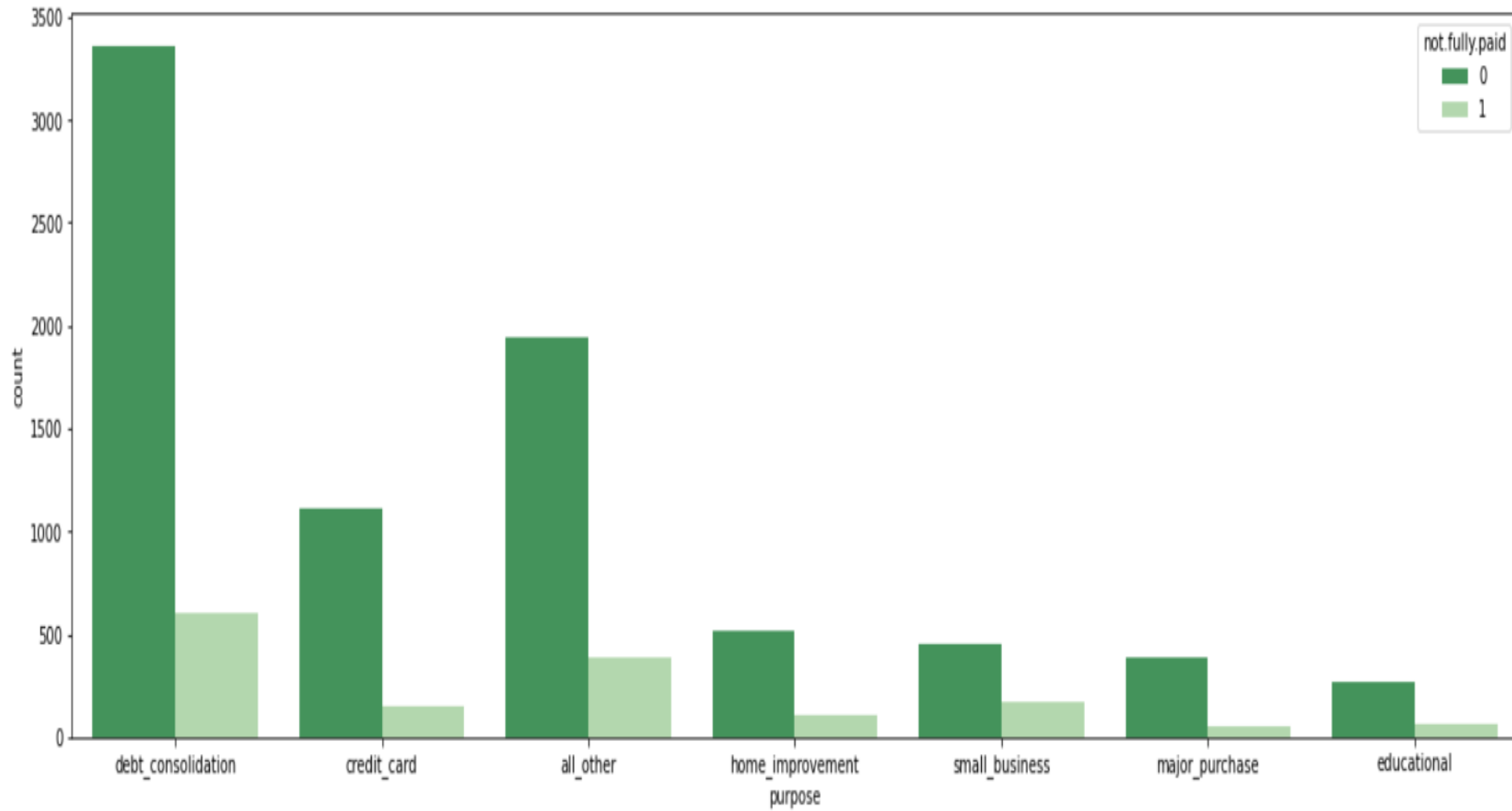
- Histogram of two **FICO** distributions on top of each other, one for each **credit.policy** outcome.



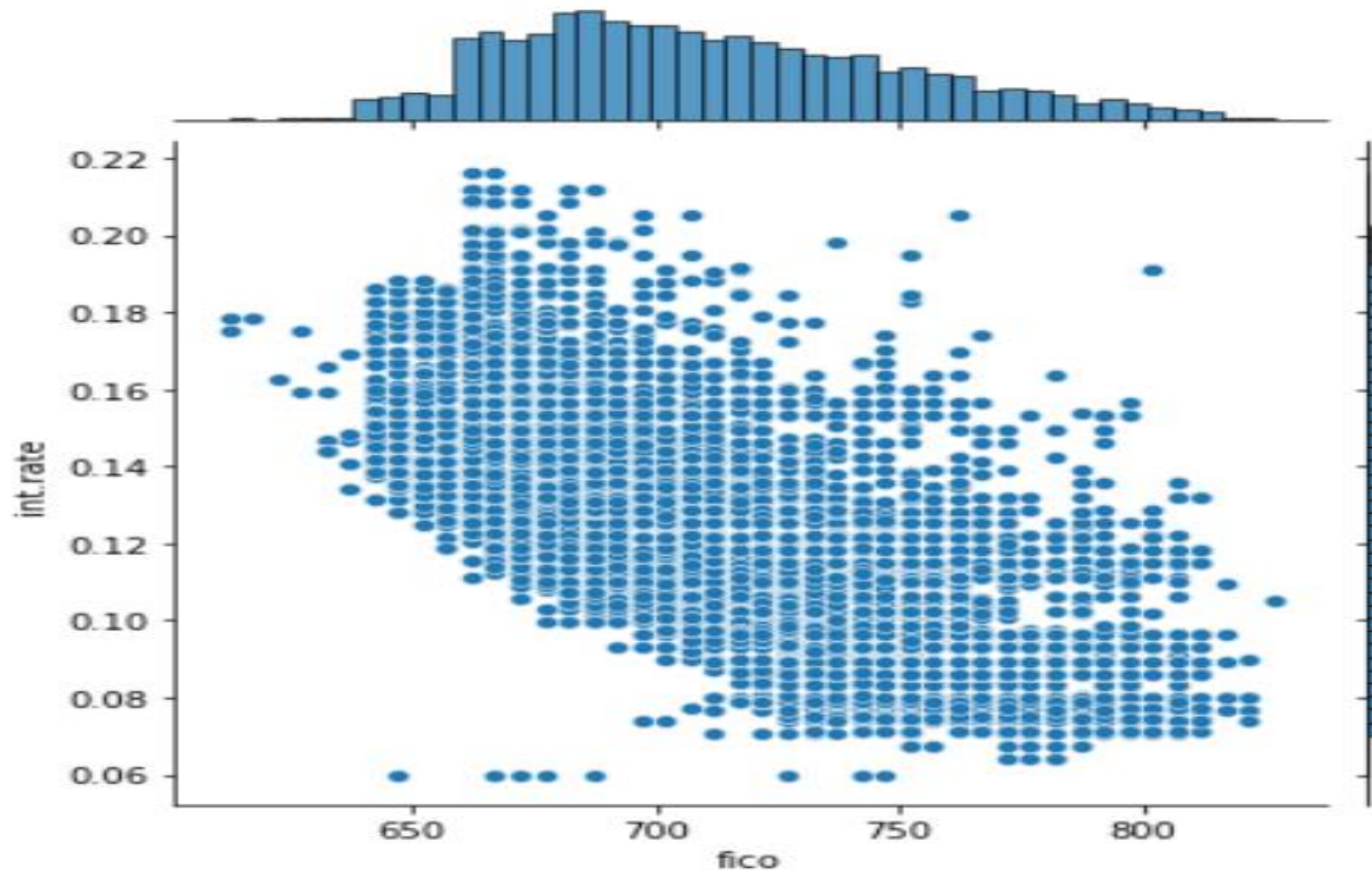
- Histogram of two FICO distributions on top of each other, one for each not.fully.paid.



- Count plot between purpose and not.fully.paid



- Join plot between FICO score and Interest rate



Observation of Graph Exploratory Analysis

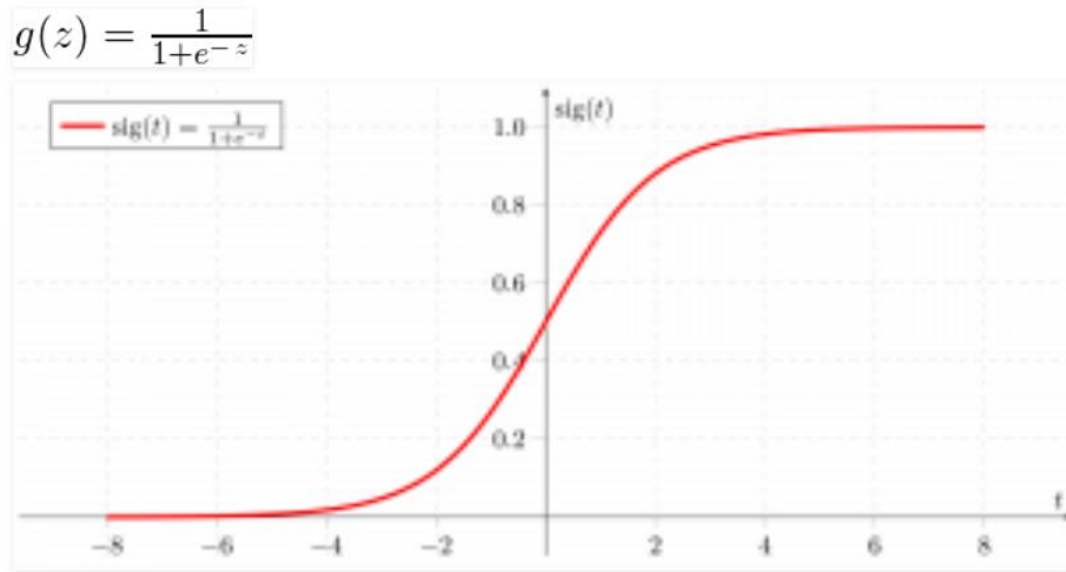
We have seen in **graph exploratory analysis** that the data distribution is not so random hence we can use both **linear** as well as **non-linear** classification model to predict the result.

We will be using **logistic regression** as our **linear** model where as in case **non-linear** model we will be using two different types of model to predict our result meticulously.

- ▶ **Tree Model - Decision tree classifier**
- ▶ **Ensemble Model - Random forest classifier**

Logistic regression

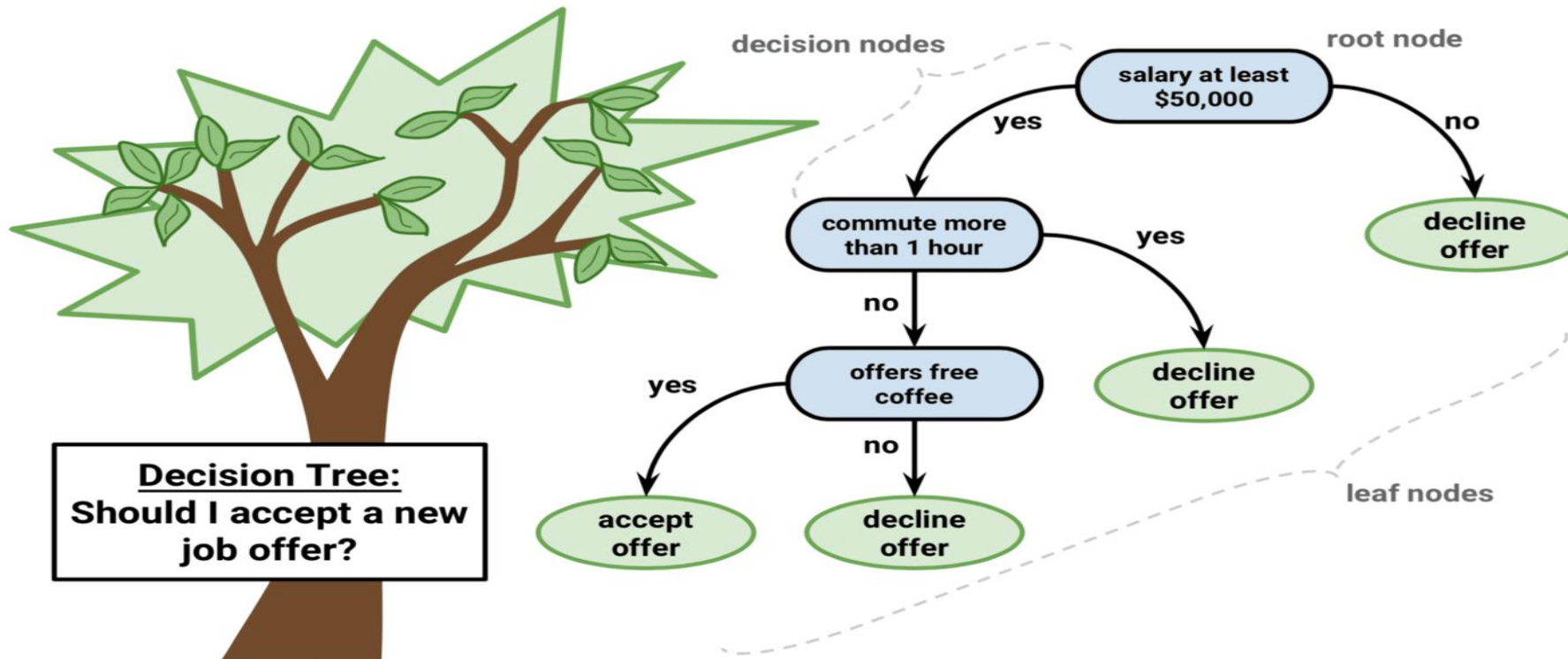
Logistic regression is a supervised learning classification algorithm used to predict the probability of a target variable. The nature of target or dependent variable is dichotomous, which means there would be only two possible classes.



In simple words, the dependent variable is binary in nature having data coded as either 1 (stands for success/yes) or 0 (stands for failure/no).

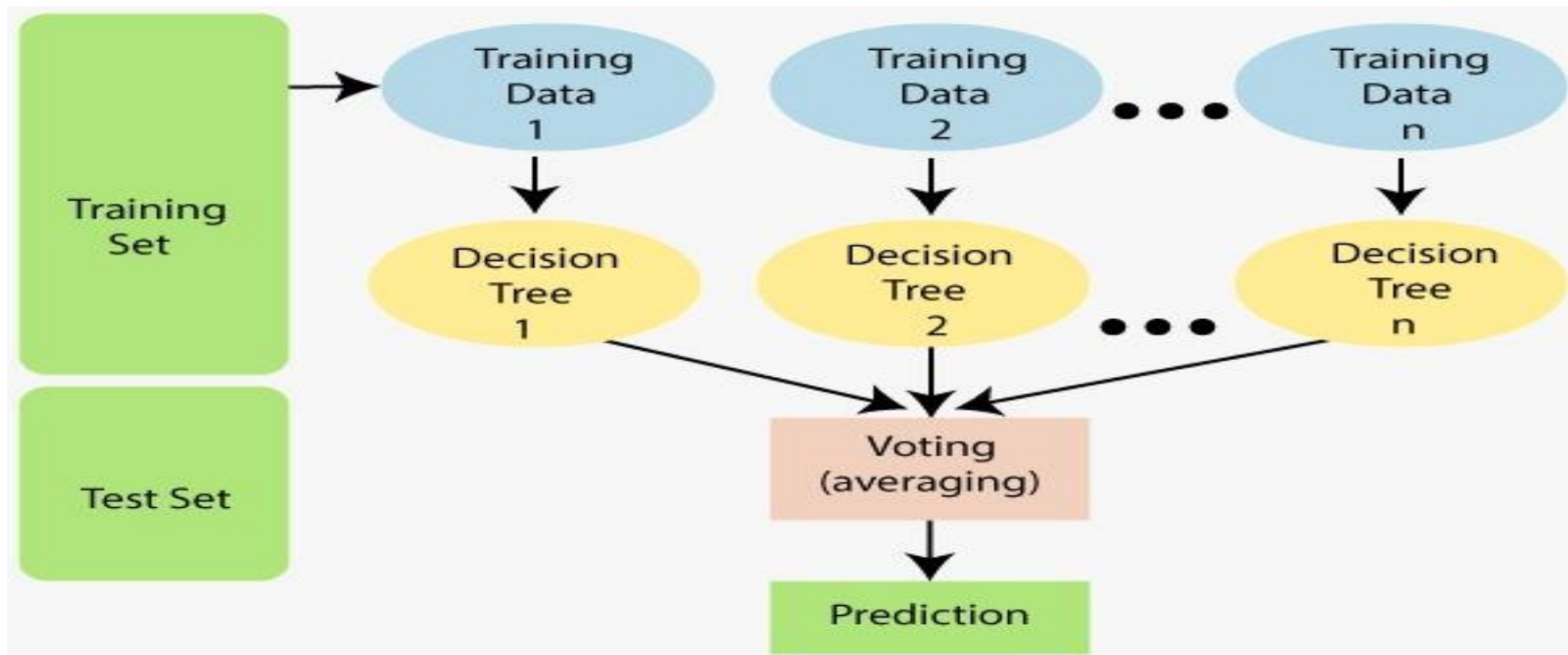
DECISION TREE CLASSIFIER

Decision tree builds classification or regression models in the form of a tree structure. It breaks down a data set into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed. The final result is a tree with decision nodes and leaf nodes. Leaf node represents a classification or decision.



RANDOM FOREST CLASSIFIER

Random forest is a supervised learning algorithm. Random forest builds multiple decision trees and merges them together to get a more accurate and stable prediction.



IMPLEMENTATION STEPS

- ▶ Importing Libraries
- ▶ Mounting Google Drive
- ▶ Importing Data
- ▶ Exploratory Data Analysis
- ▶ Encoding The String Data Column Into Integer Column
- ▶ Splitting The Dataset Into The Training Set And Test Set
- ▶ Feature Scaling
- ▶ Training The Logistic Regression Model On The Training Set
- ▶ Prediction of Result using Logistic Regression

- ▶ Building Confusion Matrix of Logistic Regression
- ▶ Accuracy of Logistic regression
- ▶ Training The Decision Tree Model On The Training Set
- ▶ Prediction of Result using Decision Tree Model
- ▶ Building Confusion Matrix of Decision Tree
- ▶ Accuracy of Decision Tree Model
- ▶ Training The Random Forest Classification Model On The Training Set
- ▶ Prediction of Result using Random Forest Model
- ▶ Building The Confusion Matrix of Random Forest Model
- ▶ Accuracy of Random Forest Model

CONCLUSION

Here is the performance of each model.

- ▶ Logistic Regression: 83.84 (false negative :379)
- ▶ Decision Tree : 73.86 (false negative : 293)
- ▶ Random Forest : 84.00 (false negative : 366)

Result: Eventually we got the maximum accuracy of **84%** with the help of the **Random Forest Model**.

[Code Link](#)

[Preview](#)

Requirements of Software and Hardware

TECHNOLOGY

- Python
- Machine Learning
- Data Science

LIBRARIES

- Numpy
- Pandas
- Sklearn
- Matplotlib
- Seaborn

SOFTWARE USED

- Microsoft Office
- Google Colab
- Window

TEAM MEMBERS

- ▶ Anil Kumar (180101005 , CSE)
- ▶ Suraj Kumar (180102040 , ECE)
- ▶ Sandeep Kumar (180101040 , CSE)
- ▶ Nikhil Kumar (180102022 , ECE)
- ▶ Karan Choudhary (180101022 , CSE)
- ▶ Ujjawal Kumar (180102043 , ECE)

A light green watercolor splash with soft, irregular edges, serving as a background for the text.

thank you