

The goal of any data mining project is to extract knowledge and patterns from data using a wide range of methods and techniques. Within big datasets, important relationships can be uncovered that give insights and can be used to make predictions that impact the business.

This project aims to develop models that are capable of predicting whether or not clients of a bank will default on a loan. For this purpose, data was collected from several bank clients, containing samples of clients who defaulted on a loan and clients who didn't.

When customers fail to make timely loan payments, banks incur losses, resulting in annual losses amounting to millions of rupees. This significantly affects the country's economic growth. In this project, we will analyse various factors, including the funded amount, location, loan balance, and more, to forecast whether an individual is likely to default on their loan.

Defaulting on a loan happens when a client misses payments for a specified period of time. When a loan defaults, it's sent to a debt collection agency, whose job it is to collect the unpaid funds from the client. The period between missing a loan payment and having the loan default is known as "delinquency." The delinquency period helps clients avoid default by giving them extra time to contact the loan servicer and catch up on missed payments.

Using the algorithms you've learned, the aim is to develop predictive models based on the data provided. The project must follow the CRISP-DM methodology and include files for its phases in the Python language, annotated in Markdown:

- data exploration and preparation;
- data pre-processing;
- creation of models using data mining algorithms;
- evaluation of the models created.

The project must be submitted with a report describing, in as much detail as possible, the process you followed to obtain your solutions. The report must include the data mining goals, the most relevant data graphical figures

and their interpretation, an explanation of the cleaning and pre-processing of data performed, an interpretation or evaluation of the models created, and commitments assumed in their development.

The work will be mainly assessed by the quality of the data analysis process, followed by the conclusions reached, policies or actions proposed, and the accuracy of the models. But more important than the accuracy of the models is the description of the analysis process and conclusions extracted from the data.

Deadline and submission instructions

- The project should be submitted to Moodle in the discipline area by **24:00 on November 19**. From this date, the note will be penalised by 10%, and no projects will be accepted after the 20th of November.
- The code and report should be placed in a ZIP file with the designation MINDD-GRPX.zip being X the group number.
- The presentation and evaluation of the project will be by group and individual on the 14th week (18–20 December).