Plan:

1. Descriptive analysis + data visualization (exploratory data analysis)
   1. Data description:
      * Problem description and research questions formulation:
        + What questions do we want to answer analysing the data
        + What potential benefits may result from the analysis? (For example, the benefit could be: a better diagnostic method, better efficiency in detecting bad/good customers applying for a loan, separating groups of customers who can be targeted with a specific offer, identifying relevant features/variables, etc.)
      * Data characteristics:
        + Data size
        + number of cases and features,
        + types of features
        + information about missing values
        + information on unusual values
          1. non-standard coding of missing values
          2. non-standard binary values
   2. Cleaning data:
      * Data type
      * Missing values as NA. How many?
      * Add missing values (several methods and choose the best)
   3. EDA:
      * Summary (min, max, quantiles, var)
      * Barplots
        + Balance
      * Histograms
        + Distribution
      * QQ Plot
      * Boxplots for int, cont
        + Outliers
        + Initial assessment of discriminative ability of consecutive features (i.e. ability to separate objects from different classes).
      * Barplots for categorises
        + Initial assessment of discriminative ability of consecutive features (i.e. ability to separate objects from different classes).
      * Histograms for categorises
        + Distribution
      * Correlations
2. Classification along with detailed accuracy assessment
   1. Methods:
      * Linear regression
      * K-NN
      * LDA
      * QDA
      * LR
      * Random tree
      * Random forest
   2. Accuracy:
      * Confusion matrix
      * Cross-Validation (CV)
        + k-fold cross-validation,
        + leave-one-out.
      * Bootstrap-based methods
        + leave-one-out bootstrap,
        + .632 estimator,
        + .632+ estimator.
      * ROC-curve