Data Mining Case study

Background information/introduction:

Being the students who have no real-world experience in data science and machine learning projects, we decided to move on with something which will enhance our data mining skills and at the same time we get to expose ourselves to the actual machine learning datasets.
The dataset we chose is an imbalanced dataset. The imbalance is in the response variable where it gets divided into a majority (94%) and minority (6%) class.

Problem Statement:

Given various features, the aim is to build a predictive model to determine the income level for people in US. The income levels are binned at below 50K and above 50K.

Data source/attributes:

The dataset has been posted on the UCI Machine learning repository here.
The key independent variables of the dataset are -

Age, Marital Status, Income, Family Members, No. of Dependents, Tax Paid, Investment (Mutual Fund, Stock), Return from Investments, Education, Spouse, Education, Nationality, Occupation, Region in US, Race, Occupation category

The dependent variable is Income and we have to predict the income range, whether the income will be less or greater than 50,000 dollars. It basically is **a Binary classification model**.

Proposed solution:

Methods to solve the problem:

1. First, we will convert the *imbalanced* into *balanced* using some sampling techniques like, under-sampling, oversampling or SMOTE analysis.
2. Then, we will apply some basic EDA and data cleaning techniques to get the underlying meaning of the predictor variables.
3. As we are done with EDA, we will get to know some features. We will try to extract those features using feature engineering.
4. Machine learning models
    a. Naïve Bayes Classifier
    b. Decision Tree
    c. Random forest
    d. SVM
    e. Xgboost

We are basically going for all the machine learning models possible and we want to compare on how these behave.

Thank you