Classification Task

In this task I get a sense of using Python to solve some classification problems.
The dataset from a Portuguese banking institution about its direct marketing campaigns.
The **goal** is to predict, if a client will subscribe a term deposit (denoted in **variable y**) or not.

1. Dataset Information

Below is the explanation of our variables from the dataset:

bank client data:
- 1 - age (numeric)
- 2 - job : type of job (categorical: 'admin.','blue-collar','entrepreneur','housemaid','management','retired','self-employed','services','student','technician','unemployed','unknown')
- 3 - marital : marital status (categorical: 'divorced','married','single','unknown'; note: 'divorced' means divorced or widowed)
- 4 - education (categorical: 'basic.4y','basic.6y','basic.9y','high.school','illiterate','professional.course','university.degree','unknown')
- 5 - default: has credit in default? (categorical: 'no','yes','unknown')
- 6 - housing: has housing loan? (categorical: 'no','yes','unknown')
- 7 - loan: has personal loan? (categorical: 'no','yes','unknown')

related with the last contact of the current campaign:
- 8 - contact: contact communication type (categorical: 'cellular','telephone')
- 9 - month: last contact month of year (categorical: 'jan', 'feb', 'mar', ..., 'nov', 'dec')
- 10 - day_of_week: last contact day of the week (categorical: 'mon','tue','wed','thu','fri')
- 11 - duration: last contact duration, in seconds (numeric). Important note: this attribute highly affects the output target (e.g., if duration=0 then y='no'). Yet, the duration is not known before a call is performed. Also, after the end of the call y is obviously known. Thus, this input should only be included for benchmark purposes and should be discarded if the intention is to have a realistic predictive model.

other attributes :
- 12 - campaign: number of contacts performed during this campaign and for this client (numeric, includes last contact)
- 13 - pdays: number of days that passed by after the client was last contacted from a previous campaign (numeric; 999 means client was not previously contacted)
- 14 - previous: number of contacts performed before this campaign and for this client (numeric)
- 15 - poutcome: outcome of the previous marketing campaign (categorical: 'failure','nonexistent','success')

social and economic context attributes from dataset with **full features**
- 16 - emp.var.rate: employment variation rate - quarterly indicator (numeric)
- 17 - cons.price.idx: consumer price index - monthly indicator (numeric)
- 18 - cons.conf.idx: consumer confidence index - monthly indicator (numeric)

- 19 - euribor3m: euribor 3 month rate - daily indicator (numeric)
- 20 - nr.employed: number of employees - quarterly indicator (numeric)

Output variable (desired target):
- 21 - y - has the client subscribed a term deposit? (binary: 'yes','no')

2. Task

The steps below served as a guidance to solve this problem.

Step 1: Data loading & Preprocessing**
- loaded the data into Python Notebook and convert it to the appropriate format (dataframe, numpy.array, list, etc.)
- observed & explored the dataset, understand each variable and its meaning
- checked for null values
- separated variables & labels

Step 2: Data Visualisation & Exploration**
- made use of learned visualisation skills to learn what is happening in your dataset
- made preliminary conclusions

Step 3: Data modelling**
- split dataset into training & testing dataset
- pick Dummy classifier as a benchmark
- fit the training dataset to the model and train the model
- output the model
- made prediction on testing dataset

Step 4: Used more advanced model**
- mapped the prediction of the testing dataset against real numbers from your dataset and compare the result
- ran Random Forest, Logistic regression and Decision Tree models

Step 5: Result extration & interpretation**
- made your conclusions and interpretation on the model and final results
- evaluated the performance of the model and algorithm

3. Findings

The best model suitable is the Decision Tree as it has the highest recall score.