Week 7

Healthcare - Persistency of a drug: Group Project

Name: Judith Chepngetich

Email: chepngetich.judith@gmail.com

Group Name: N/A-(Self)

Country: Kenya

Company: Kenya Revenue Authority

Specialization: Data Science

Github Repo Link: https://github.com/JudieChep/Assignments.git

Problem Description

ABC pharma Company is in the pharmaceutical business and has a challenge in determining the persistence of a drug. The company would like to automate the process of identification of persistency since there are several factors that need to be considered to determine this. An insight into which factors affect persistency will be useful in automating this process.

Business Understanding

Business Process

Being a pharmcaeutical company , the company manufactures drugs to be used by healthcare professionals and patients to improve healthcare.The process is depicted in the below diagram:

Problems/Challenges

ABC ompany has a challenge identifying the factors that affect /influence persistency in patients. Data on clinical, demographic and disease treatment are available and there is need to automate the process so as to bring efficiency.

Business Objective

The business objective is to maintain high persistent levels in order to promote health among the patients, which will eventually bring higher revenue to the company.

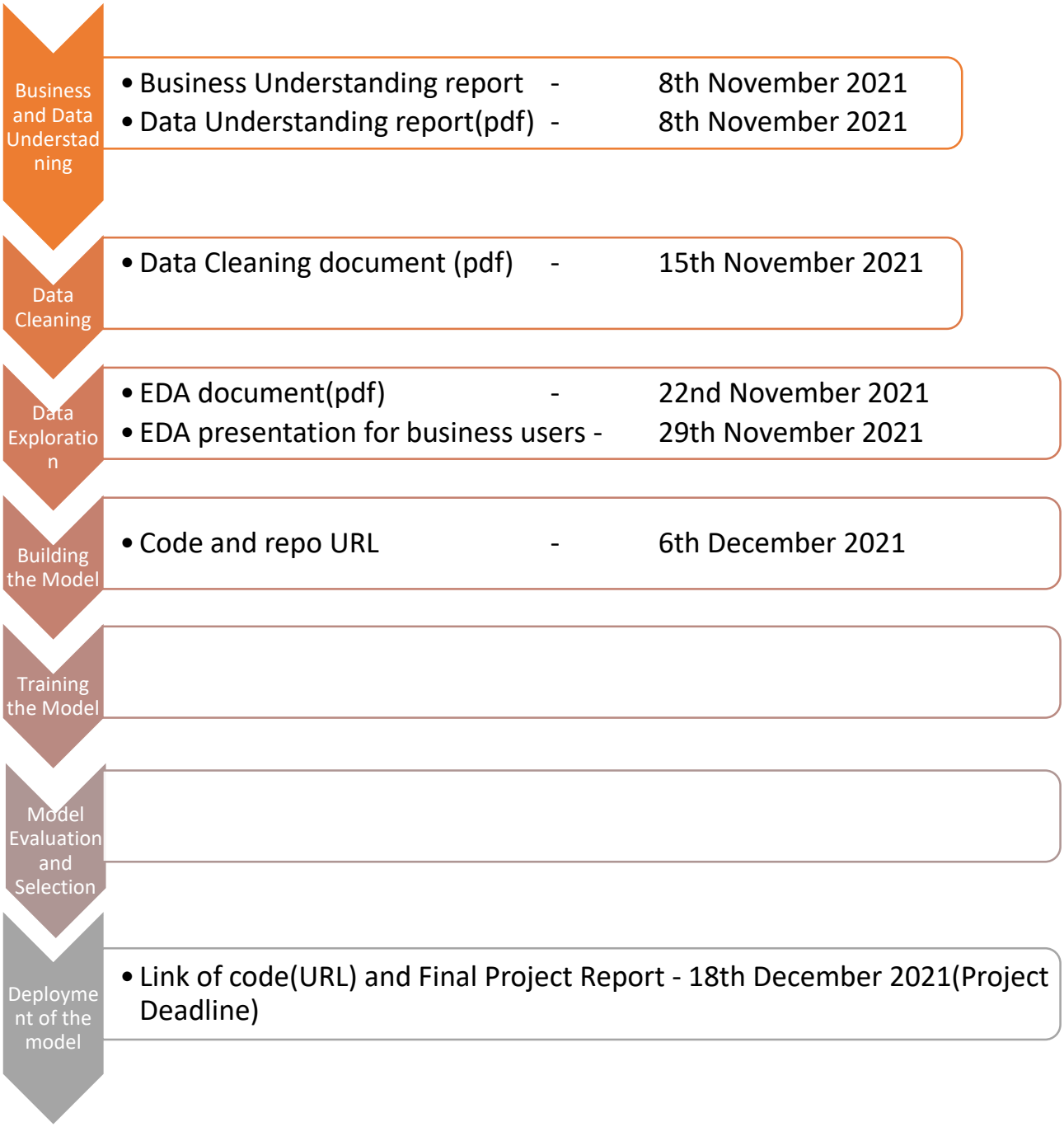
Successs Criteria

Enhanced/improved persistence rate in patients.

Project Lifecycle

The below diagram illustrates the project lifecycle that will be adopted for this project:

Activities	Deliverables


Business and Data Understanding
• Business Understanding report - 8th November 2021
• Data Understanding report(pdf) - 8th November 2021
Data Cleaning
• Data Cleaning document (pdf) - 15th November 2021
Data Exploration
• EDA document(pdf) - 22nd November 2021
• EDA presentation for business users - 29th November 2021
Building the Model
• Code and repo URL - 6th December 2021
Training the Model
Model Evaluation and Selection
Deployment of the model
• Link of code(URL) and Final Project Report - 18th December 2021(Project Deadline)

Project Activities

1. *Business and Data Understanding*- A description of the business process, objective and its challenges.Data will be accessed/downloaded from the Google Drive link provided.
2. *Data cleaning*- The pandas profiling report will be used to understand the data and check for any missing values or outliers within the dataset. Any categorical variables that need encoding will also be processed at this stage. The missing values will be imputed using either mean, mode or median .
3. *Data Exploration*-Analysis shall be conducted on the data to identify whether some variables are more important in determining persistency compared to others.
4. *Building model* –Three model candidates will be built using Support Vector Classifier,K Neighbours Classifier and Naives Bayes classification algorithms.
5. *Training of the model*-The data will be split into training and test set using stratified sampling from *sklearn* library.
6. *Model Evaluation and selection* -The model will be evaluated using accuracy, precision and recall . The ROC AUC will be also be determined, and the model with better performance will be selected as the main optimal model that will be deployed for use.
7. *Deployment of model*- The model will be saved in pickle and a Flask web service used for deployment