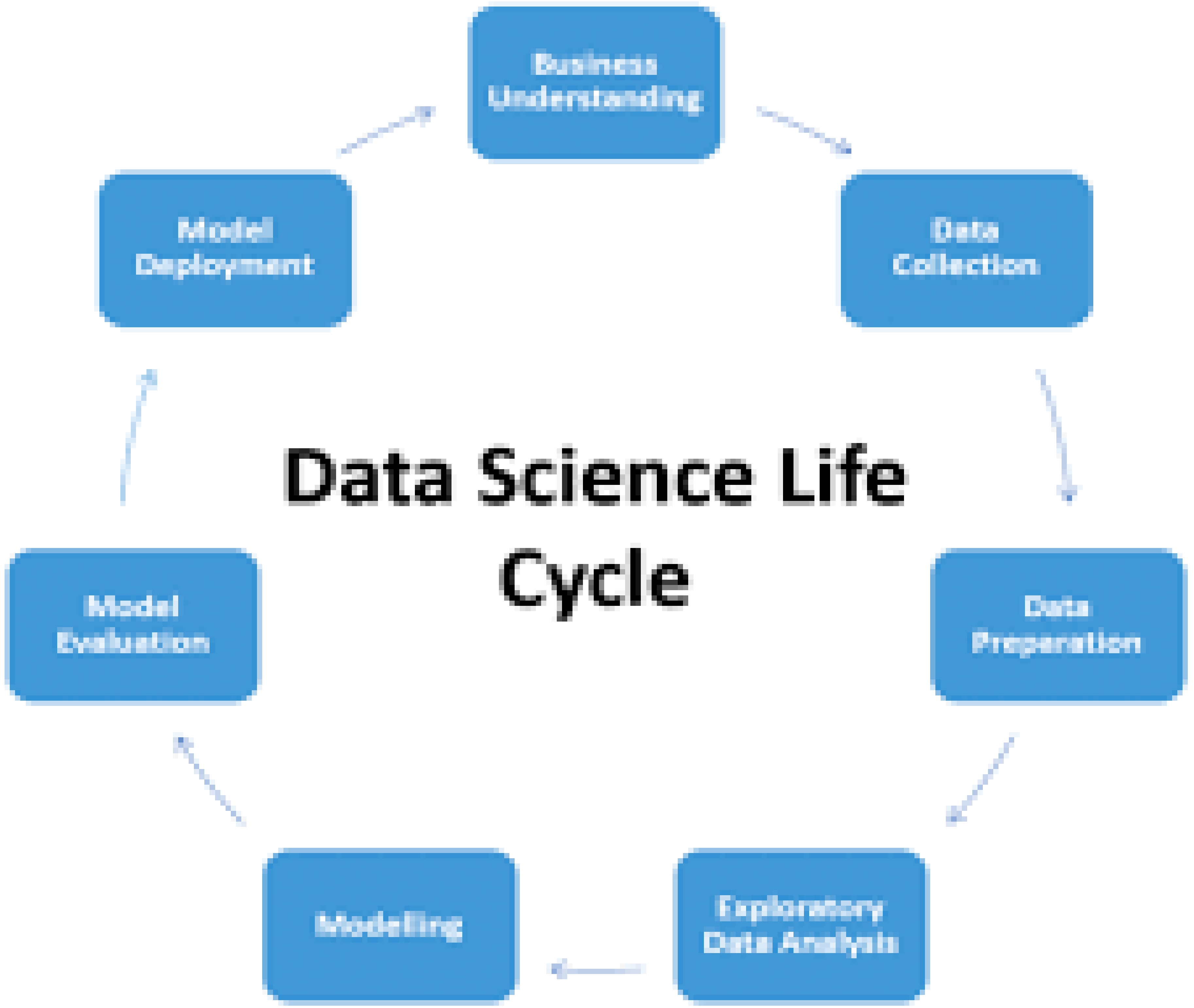
Problem Statement: Health Insurance Lead Prediction

Venkatesh Kalyane

Approach to solve the given problem



Information regarding given Dataset

- Given Problem statement is a classification problem and the Target Variable is Categorical Variable
- Presence of both Continuous Variables
- Presence of Null Values
- Presence of Highly correlated Variables
- Presence of Outliers
- Dataset is imbalanced

Selection of Model by considering the given dataset

Random Forest Classifier

- It can handle both Continuous and Categorical Variables and Less likely to overfit
- Robust to Outliers
- Feature Scaling is not required
- Require Less Parameter tuning
- Imbalance can be handled using class_weight parameter

Cleaning and Handling the Data

- Missing Values - Replaced Numeric values with mean based on the data distribution and applied backward filling method for Non-Numeric values
- Encoding- Create Dummy Varibles for Categorical features and apply label encoding
- Checking the presence of Duplicated values
- Remove the Highly Correlated Variables
- Tune the Hyperparameters

Github Link for Solution

https://github.com/Venkatesh-Kalyane/Health-Insurance-Lead-Prediction