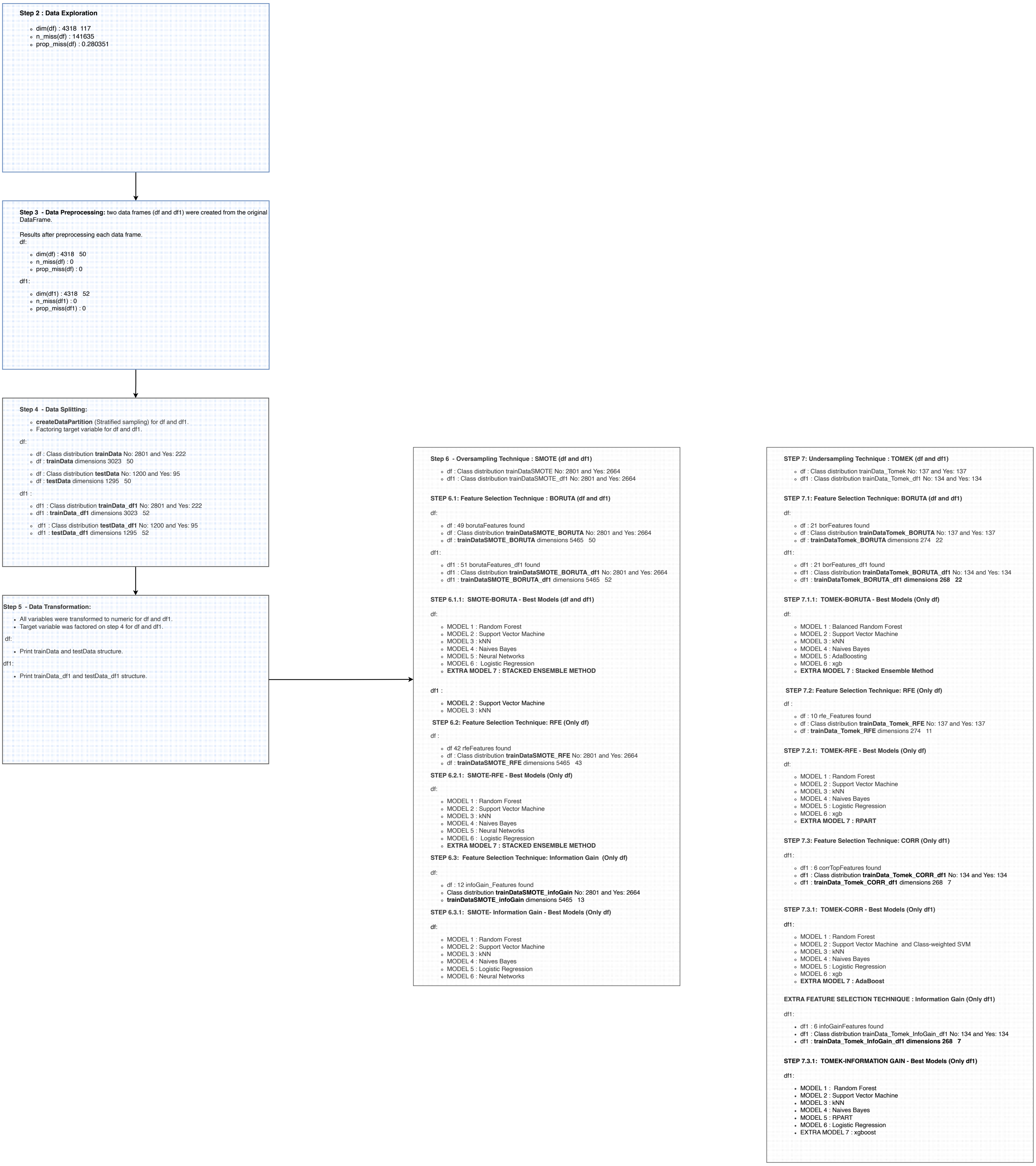

Step 2 : Data Exploration

- dim(df) : 4318 117
- n_miss(df) : 141635
- prop_miss(df) : 0.280351

Step 3 - Data Preprocessing: two data frames (df and df1) were created from the original DataFrame.

Results after preprocessing each data frame.
df:

- dim(df) : 4318 50
- n_miss(df) : 0
- prop_miss(df) : 0

df1:

- dim(df1) : 4318 52
- n_miss(df1) : 0
- prop_miss(df1) : 0

Step 4 - Data Splitting:

- **createDataPartition** (Stratified sampling) for df and df1.
- Factoring target variable for df and df1.

df:

- df : Class distribution **trainData** No: 2801 and Yes: 222
- df : **trainData** dimensions 3023 50

- df : Class distribution **testData** No: 1200 and Yes: 95
- df : **testData** dimensions 1295 50

df1 :

- df1 : Class distribution **trainData_df1** No: 2801 and Yes: 222
- df1 : **trainData_df1** dimensions 3023 52

- df1 : Class distribution **testData_df1** No: 1200 and Yes: 95
- df1 : **testData_df1** dimensions 1295 52

Step 5 - Data Transformation:

- All variables were transformed to numeric for df and df1.
- Target variable was factored on step 4 for df and df1.

df:

- Print trainData and testData structure.

df1:

- Print trainData_df1 and testData_df1 structure.

Step 6 - Oversampling Technique : SMOTE (df and df1)

- df : Class distribution trainDataSMOTE No: 2801 and Yes: 2664
- df1 : Class distribution trainDataSMOTE_df1 No: 2801 and Yes: 2664

STEP 6.1: Feature Selection Technique : BORUTA (df and df1)

df:

- df : 49 borutaFeatures found
- df : Class distribution **trainDataSMOTE_BORUTA** No: 2801 and Yes: 2664
- df : **trainDataSMOTE_BORUTA** dimensions 5465 50

df1:

- df1 : 51 borutaFeatures_df1 found
- df1 : Class distribution **trainDataSMOTE_BORUTA_df1** No: 2801 and Yes: 2664
- df1 : **trainDataSMOTE_BORUTA_df1** dimensions 5465 52

STEP 6.1.1: SMOTE-BORUTA - Best Models (df and df1)

df:

- MODEL 1 : Random Forest
- MODEL 2 : Support Vector Machine
- MODEL 3 : kNN
- MODEL 4 : Naives Bayes
- MODEL 5 : Neural Networks
- MODEL 6 : Logistic Regression
- **EXTRA MODEL 7 : STACKED ENSEMBLE METHOD**

df1 :

- MODEL 2 : Support Vector Machine
- MODEL 3 : kNN

STEP 6.2: Feature Selection Technique: RFE (Only df)

df :

- df 42 rfeFeatures found
- df : Class distribution **trainDataSMOTE_RFE** No: 2801 and Yes: 2664
- df : **trainDataSMOTE_RFE** dimensions 5465 43

STEP 6.2.1: SMOTE-RFE - Best Models (Only df)

df:

- MODEL 1 : Random Forest
- MODEL 2 : Support Vector Machine
- MODEL 3 : kNN
- MODEL 4 : Naives Bayes
- MODEL 5 : Neural Networks
- MODEL 6 : Logistic Regression
- **EXTRA MODEL 7 : STACKED ENSEMBLE METHOD**

STEP 6.3: Feature Selection Technique: Information Gain (Only df)

df:

- df : 12 infoGain_Features found
- Class distribution **trainDataSMOTE_infoGain** No: 2801 and Yes: 2664
- **trainDataSMOTE_infoGain** dimensions 5465 13

STEP 6.3.1: SMOTE- Information Gain - Best Models (Only df)

df:

- MODEL 1 : Random Forest
- MODEL 2 : Support Vector Machine
- MODEL 3 : kNN
- MODEL 4 : Naives Bayes
- MODEL 5 : Logistic Regression
- MODEL 6 : Neural Networks

STEP 7: Undersampling Technique : TOMEK (df and df1)

- df : Class distribution trainData_Tomek No: 137 and Yes: 137
- df1 : Class distribution trainData_Tomek_df1 No: 134 and Yes: 134

STEP 7.1: Feature Selection Technique: BORUTA (df and df1)

df:

- df : 21 borFeatures found
- df : Class distribution **trainDataTomek_BORUTA** No: 137 and Yes: 137
- df : **trainDataTomek_BORUTA** dimensions 274 22

df1:

- df1 : 21 borFeatures_df1 found
- df1 : Class distribution **trainDataTomek_BORUTA_df1** No: 134 and Yes: 134
- df1 : **trainDataTomek_BORUTA_df1 dimensions 268 22**

STEP 7.1.1: TOMEK-BORUTA - Best Models (Only df)

df:

- MODEL 1 : Balanced Random Forest
- MODEL 2 : Support Vector Machine
- MODEL 3 : kNN
- MODEL 4 : Naives Bayes
- MODEL 5 : AdaBoosting
- MODEL 6 : xgb
- **EXTRA MODEL 7 : Stacked Ensemble Method**

STEP 7.2: Feature Selection Technique: RFE (Only df)

df :

- df : 10 rfe_Features found
- df : Class distribution **trainData_Tomek_RFE** No: 137 and Yes: 137
- df : **trainData_Tomek_RFE** dimensions 274 11

STEP 7.2.1: TOMEK-RFE - Best Models (Only df)

df:

- MODEL 1 : Random Forest
- MODEL 2 : Support Vector Machine
- MODEL 3 : kNN
- MODEL 4 : Naives Bayes
- MODEL 5 : Logistic Regression
- MODEL 6 : xgb
- **EXTRA MODEL 7 : RPART**

STEP 7.3: Feature Selection Technique: CORR (Only df1)

df1:

- df1 : 6 corrTopFeatures found
- df1 : Class distribution **trainData_Tomek_CORR_df1** No: 134 and Yes: 134
- df1 : **trainData_Tomek_CORR_df1** dimensions 268 7

STEP 7.3.1: TOMEK-CORR - Best Models (Only df1)

df1:

- MODEL 1 : Random Forest
- MODEL 2 : Support Vector Machine and Class-weighted SVM
- MODEL 3 : kNN
- MODEL 4 : Naives Bayes
- MODEL 5 : Logistic Regression
- MODEL 6 : xgb
- **EXTRA MODEL 7 : AdaBoost**

EXTRA FEATURE SELECTION TECHNIQUE : Information Gain (Only df1)

df1:

- df1 : 6 infoGainFeatures found
- df1 : Class distribution trainData_Tomek_InfoGain_df1 No: 134 and Yes: 134
- df1 : **trainData_Tomek_InfoGain_df1 dimensions 268 7**

STEP 7.3.1: TOMEK-INFORMATION GAIN - Best Models (Only df1)

df1:

- MODEL 1 : Random Forest
- MODEL 2 : Support Vector Machine
- MODEL 3 : kNN
- MODEL 4 : Naives Bayes
- MODEL 5 : RPART
- MODEL 6 : Logistic Regression
- EXTRA MODEL 7 : xgboost