Proposed Methodology:

The proposed method, preprocessing steps and the experiment design is described below.

Architecture of proposed method : -

The proposed PNN based imputation technique is shown in the figure.

Let X denote the dataset and C denote the number of attributes. Let XD denote the dataset with all the values of X converted in to nominal numbers. For the experiment purpose 5% data of XD is replaced by ‘?’. Let XCR denote the set of complete records and XIR denote the set of incomplete records. All the rows of XD that does not contain ‘?’ is appended to XCR and the rest are appended to XIR. For XIR the mode of each attribute is calculated and stored in a list. The mode is calculated by ignoring the missing values in each attribute.

The resulting dataset is termed as XIM. Taking attribute 1 as the output and all other attributes as the input dataset X1 is generated removing the 1st attribute from XIM.

Now a list X\_true is made which contains all the correct values at the position of missing values for 1st attribute. Dataset X\_M and XTR are made from X1 and XCR respectively taking 1st attribute as the output. One Hot Encoding is applied on X\_M and XTR and they are converted into indicator matrices. Dataset Xa and Xb are created which contains all the rows of XCR with nominal value 1 and 2 respectively in 1st attribute. Now 10 fold cross validation is used . In each fold a test set is generated by taking the rows from X\_M at the positions which contain ‘?’ in 1st column of XIR. Train set for class 1 of 1st attribute is Xa and that of class 2 is Xb. Now PNN is used to classify each row of X\_test into class 1 or 2 and is stored in list all\_pred\_values. For each fold percentage correct prediction (PCP) is calculated by comparing the predicted class with the true class . These steps are then repeated for all the other attributes.

**Algorithm for PNN based data imputation**

**{**

for Xij in XD

if Xij is missing then

Add Xi to XIR

Else

Add Xi to XCR

for each j in (0, number of attributes)

modej = Calc\_Mode(XD,j)

for each i in rows

for each j in column

if Xij is missing then

Xij=modej

Add Xi to XIM

for i in rows of XIR

if XIR[i][0] is missing then

add ith row of XIR to X1

for each i in rows

for each j in column

if X1ij is missing then

X1ij=modej

for i in rows of XD

if XD[i][0] is missing then

add X\_complete[i][0] to X\_true

for i in rows of X1

Add X1[i][1:6] to XPR

for i in rows of XCR

Add XCR[:][1:6] to XTR

Convert XTR into indicator matrix

for i in rows of XIR

Add XIR[:][1:6] to X\_M

for each i in rows

for each j in column

if X\_Mij is missing then

X\_Mij=modej

Convert X\_M to indicator matrix

For i in rows of XCR

If XCR[i][0] is 1 then

Add XCR[i][:] to Xa

If XCR[i][0] is 2 then

Add XCR[i][:] to Xb

Train PNN using X\_M, Xa and Xb with SF as alpha

Now predict the values for X\_M

}

**Experimental Design:-**

The dataset used here is Process Safety Dataset. The experiment is carried out with 5% missing data. Since in the original dataset no data was missing , we removed 5% of the total data randomly. All the records with missing values was removed from the dataset and the rest was used for experimenting purpose. To get the best result 10 folds cross validation was used. ith test fold was tested by training the dataset on XTR. To check the accuracy of the prediction Percentage Correct Prediction(PCP) is calculated for each test fold.