df = pd.read\_csv(‘file.csv’)

#Identify duplicate values

df = df.duplicated()

#Removing duplicates

df\_unique = df.drop\_duplicates()

Dimensionality Reduction

Dimensionality reduction is the process of reducing the number of random variables under consideration, by obtaining a set of principal variables. It can be divided into (1)feature selection and (2)feature extraction.

There are two components of dimensionality reduction:

* **Feature selection:** In this, we try to find a subset of the original set of variables, or features, to get a smaller subset which can be used to model the problem. It usually involves three ways:
  1. Backward Elimination
  2. Forward Selection
  3. Bidirectional Elimination
  4. Score Comparison

Backward Elimination

Step 1: Select a significance level to stay in the model. (eg SL=0.05)

Step 2: Fill the full model with all possible predictors.

Step 3: Consider the predictor with the highest p-value. If p>SL, goto step 4 else finish.

Step 4: Remove the predictor.

Step 5: Fit model without the variable.

Forward Selection

Step 1: Select a significance level to stay in the model. (eg SL=0.05)

Step 2: Fit all simple simple regression models. Select the one with lowest p-value.

Step 3: Keep this variable and fit all possible models with one extra predictor added to the one you already have.

Step 4: Consider the predictor with the highest p-value. If p>SL, goto step 4 else finish.

. Bidirectional Elimination

Step 1: Select a significance level to enter and stay in the model. (eg SLENTER=0.05, SLSTAY=0.5)

Step 2: Perform the next step of Forward Selection(new variables must have p<SLENTER to enter).

Step 3: Perform all step of Backward Elimination(old variables must have p<SLSTAY to stay).

Step 4: No new variables can enter and no old variables can exit.

* **Feature extraction:** This reduces the data in a high dimensional space to a lower dimension space, i.e. a space with lesser no. of dimensions.
  1. PCA(Principal Component Analysis)
  2. LDA(Linear Discriminant Analysis)
  3. Kernel PCA

PCA(Principal Component Analysis)

Goal:

* Identify patterns in data.
* Detect the correrlation between variables.
* Reduce the dimension of a d-dimensional dataset by projecting it onto a k- dimensional subspace

Operation:

* Standardize the data.
* Obtain the eigen vectors and eigen values from the covariance matrix or correlation matrix, or perform Singular Vector decomposition.
* Sort the eigen value in descending order and choose the k eigen vectors that corresponds to the k largest eigen values where k is the number of dimensions o the new feature subspace(k<=d)
* Construct the projection matrix W from the selected k eigen vectors.
* Transform the original dataset X via W to obtain a k dimensional feature subspace Y.
* From the m independent variables of your dataset PCA extracts p<=m new independent variables that explain the most of the variance of the dataset REGARDLESS OF THE DEPENDENT VARIABLE.

The fact that dependent variable is not considered makes PCA an unsupervised model.

LDA(Linear Discriminant Analysis)

LDA differs because in addition to finding the component axes with LDA we are interested in the axes that maximize the separation between multiple classes.

Operation:

* Compute the d-dimensional mean vectors for the different classes from the dataset
* Compute the scatter matrices.
* Compute the eigen vectors and corresponding eigen values for the scatter matrix
* Sort the eigen vectors by decreasing eigen values and choose the k eigen vectors with the largest eigenvalues to form a d X k dimensional matrix W.
* Use this d X k eigenvector matrix to transform the samples onto the new subspace
* From the n independent variables of your dataset LDA extracts p<=n new independent variables that separates the most of the classes of the DEPENDENT VARIABLE.

The fact that dependent variable is considered makes LDA a supervised model.