APPLYING FEATURE SELECTION TECHNIQUES ON HIGH DIMENSIONAL DATA TO EXTRACT BEST FEATURES

- Dimensionality referred to as number of input variables or features for a given data set.
- Dimensionality reduction refers to techniques that reduce the number of input variables in a data set.
- High dimensionality statistics and dimensionality reduction techniques are often used for data visualization. These techniques can be used in applied machine learning to simplify a classification or regression dataset in order to better fit a predictive model.

Problem with high dimensions:

- Performance of machine learning algorithms may degrade with too many input variables.
- If data is represented in rows and columns. Columns are the input variables that are fed as input to model to predict target variable. Input variables are also called as features. Having large number of dimensions in feature space means volume of space is very large and often the rows of data may be small sample. This can impact the performance of model to fit on data with many input features. Therefore it is desirable to reduce number of input features.

Dimensionality Reduction:

- Fundamental reason for curse of dimensionality is that high dimensional functions have the potential to be much more complicated than low dimensional ones and those complications are harder to find. The only way to solve this by incorporating knowledge about the data that is correct.
- Dimensioanlity reduction is a data preparation technique performed data prior to modelling. It might be performed after data cleaning and data scaling and before training a predictive model.

Feature selection methods:

- These are most common techniques that use scoring or statistical methods to select which features to keep and which features to delete.
- We perform feature selection to remove irrelevant features that donot help much with the classification problem.
- Feature selection is also called as attribute selection or variable selection.
- Feature selection is different from dimensionality reduction. Both methods try to reduce the number of attributes in the data set but dimensionality reduction method creates new combinations of attributes present in the data without changing them.
- The objectives of feature selection are:
  - Improving the prediction performance of the predictors.
  - Providing faster and more cost-effective predictors.
  - Providing a better understanding of the underlying process that generated the data.

Methodology:

- The data set used in the project is a sample data set of chronic obstructive pulmonary disease (COPD) prediction with 600 rows and 320 columns.
- The sample data set is a high dimensional dataset. When this data is used by a classification models the accuracy obtained is 51.33%.

- As a preprocessing step we have applied resampling on the data.Resampling is the method that consists of drawing repeated samples from the original data samples.Resampling generates a unique sampling distribution on the basis of the actual data. The method of resampling uses experimental methods, rather than analytical methods, to generate the unique sampling distribution.
- After the application of resampling and then using the data set for the classification model the accuracy is 82.5%.
- We have applied info gain,gain ratio and correlation attribute techniques of feature selection to extract the best attributes.
- Info gain is the most commonly used feature selection technique. It is one of the filter methods,it evaluates the gain of each variable in the context of the target variable. It uses ranker search method,in which ranking is given to the variables and these ranks helps to decide the features that need to be removed. Information gain calculates the statistical dependence between two variables.
- Gain ratio evaluates the worth of an attribute by measuring the gain ratio with respect to the target class.Gain ratio is an extension to IG measure.This measure overcomes bias of IG toward the features with the large number of values by applying normalization to information gain.
- Correlation method is used to measure the correlation between each of the attributes and the target class attribute. It considers nominal attributes in value bias and each value acts as an indicator. The combination of the attribute evaluator along with the rank search method is applied on the data set. Correlation is a statistical term that explains how close two variables are to have a linear relationship with each other. The correlation coefficient lies between -1 to +1. features with high correlation are more linearly dependent and hence have the same effect on the dependent variable. So when two features have high correlation, we can drop one of the two features.
- After applying the feature selection techniques and considering the top 250,top 150 and top 100 features of data for modelling it is observed that the top 100 attributes obtained from gain ratio and top 150 attributes obtained from correlation attribute evaluator have shown accuracy **83%**.
- We again apply the resampling on the top 100 and top 50 attributes and then passed them as input attributes for modeling of classification models,we obtained that the resampled top 100 attributes of Info gain and Gain ratio have shown an accuracy of **90.83%**.
- So these attributes can be referred as the best features that are extracted from the high dimensional data which helps in providing the effective performance measures of the model.