HOMEWORK ASSIGNMENT-3

"Using predictive analytics and big data to optimize pharmaceutical outcomes"

In healthcare, the data is collected in large quantities from various devices, social media, smartphone applications, and so on which is known as big data. Big data refers to complex and large datasets consisting of different types of information, which collect data from different sources at different times and distances. In terms of healthcare, the data is collected from data processing applications along with data which is collected from smartphone applications, social media, and so on. In the recent literature, big data is referred to as the data sets in which product sample size and number of variables are greater than log 7. So, therefore the three main characteristics of big data are large sample size, high heterogeneity, and high dimensionality.

Challenges of big data- Due to its high dimensionality and large sample, big data is associated with some problems like noise accumulation and spurious correlation. Noise accumulation is referred to as the accumulation of estimation errors when the prediction is based on a large number of parameters, which leads to poor classification or poor prediction, which in turn means that the addition of more predictors will not improve the predictive power but it leads to accumulated noise. Spurious correlation refers to the fact that when a large number of variables are evaluated, important variables can be highly correlated with variables with which they have no actual relations. The analytical challenges often make the traditional statistical methods invalid in analyzing big data. The new methods of analysis developed to overcome these challenges are often referred to as predictive analytics. The application of predictive

analytics is in many fields like actuarial science, marketing, finance, retailing, etc. The first step in order to apply predictive analytics in healthcare is to build up a unique dataset which is often constructed at the patient level. These are the following steps in order to construct the dataset – 1) cleansing and normalizing the original dataset in order to let the information constant across different data sources.

2) then the aggregation of those data sources into a single dataset at the patient level. 3) deidentification of the protected information according to Health Insurance portability. 4) the last step means validation of the process to ensure the accuracy of the data.

Preparation of input dataset- Reduction of dimensionality- Although the predictive analytic algorithms can handle a large number of input predictors, the procedure of reducing the number of input dimensions can improve the performance of models with high dimensionality. The assumption on which the dimensionality reduction method is based on high dimensional data set lies on or near a lower-dimensional manifold. In some cases, dimensionality reduction techniques can be applied to transform the original high-dimensional data set into a compact lower-dimensionality expression.

Steps involved in the development of big data-driven predictive analytics- I) The data is aggregated into a unified dataset from different data sources at the patient level. II) input dataset for the study sample is generated by identifying the study sample including observations and variables of interest. III) dimensionality reduction technique is applied. IV) then the dataset is randomly split into training, validation, and testing data. V) from which the training dataset is used for model building with multivariate variable selection along with

univariate variable selection. VI) the validation dataset is used for calculating performance measures for all models in order to compare them and then selecting the final model on the basis of the calculated performance measure. VII) then finally the test dataset is used for calculating and evaluating the performance of the final model.

In this whole process, the blue boxes are representing the datasets, the red boxes represent the processes, the yellow stars represent the models which are constructed, and the purple ones represent the models selected. And after this whole process of dimensionality reduction, still, some predictors are left so in order to overcome that, the elimination of useless covariates can be performed. The disadvantage of predictive analysis is their tendency to overfit. That is caused due to complex models and the presence of some noise in the underlying model. So, to overcome this overfitting problem, the dataset is split into training, validation, and testing data. Models commonly used in big data analytics include artificial neuronal networks, support vector machines, discriminant analysis, and classification trees. Machine learning models are often referred to as “black box” approaches, which means that they produce predictions but do not provide an understanding of how they do so. In the development of multivariable models in predictive analytics, it is common to build several models using different algorithms or different parameters and compare their performance. Depending on the types of models to be developed, a computer scientist with some capability in programming languages and advanced predictive analytics may be needed to conduct advanced predictive modeling. The use of predictive analytics in optimizing medication will likely transform current clinical practice.

So finally, we can conclude that predictive analysis can become a tool to improve patient outcomes and leverage big data as well.