**Survival Analysis - Cancer Prediction**

**Introduction**

Over the past two decades, there has been an explosion in the use of digital footprints to monitor and predict human behaviors. The source of data used for this purpose is our online use of the internet, the emails we send and transactions we make. Analysis of these footprints through machine-learning techniques (MLT) has been exploited in the public domain by government and business to predict behaviors and inform investment decisions. In research, MLT have also been used to analyze gene expression dataand for medical image analysis However to date, there has been little exploration of these methodologies in the clinical setting. We hypothesized that MLT may offer a paradigm shift in clinical medicine that can address core issues with large and complex data sets. These techniques offer the potential to derive adaptive systems from diverse data sets, discover latent connections between data items and to predict outcomes.

**Objective of Research**

Using the prediction of cancer outcome as a model, we have tested the hypothesis that through analyzing routinely collected digital data contained in an electronic administrative record (EAR), using machine-learning techniques, we could enhance conventional methods in predicting clinical outcomes.

**Problem Statement**

The task of this project is to predict a survival analysis of cancer suffering patients. Based on the factors like age, operation\_years, axillary nodes the prediction process is depend on and proportional to prediction.

**Industry Profile**

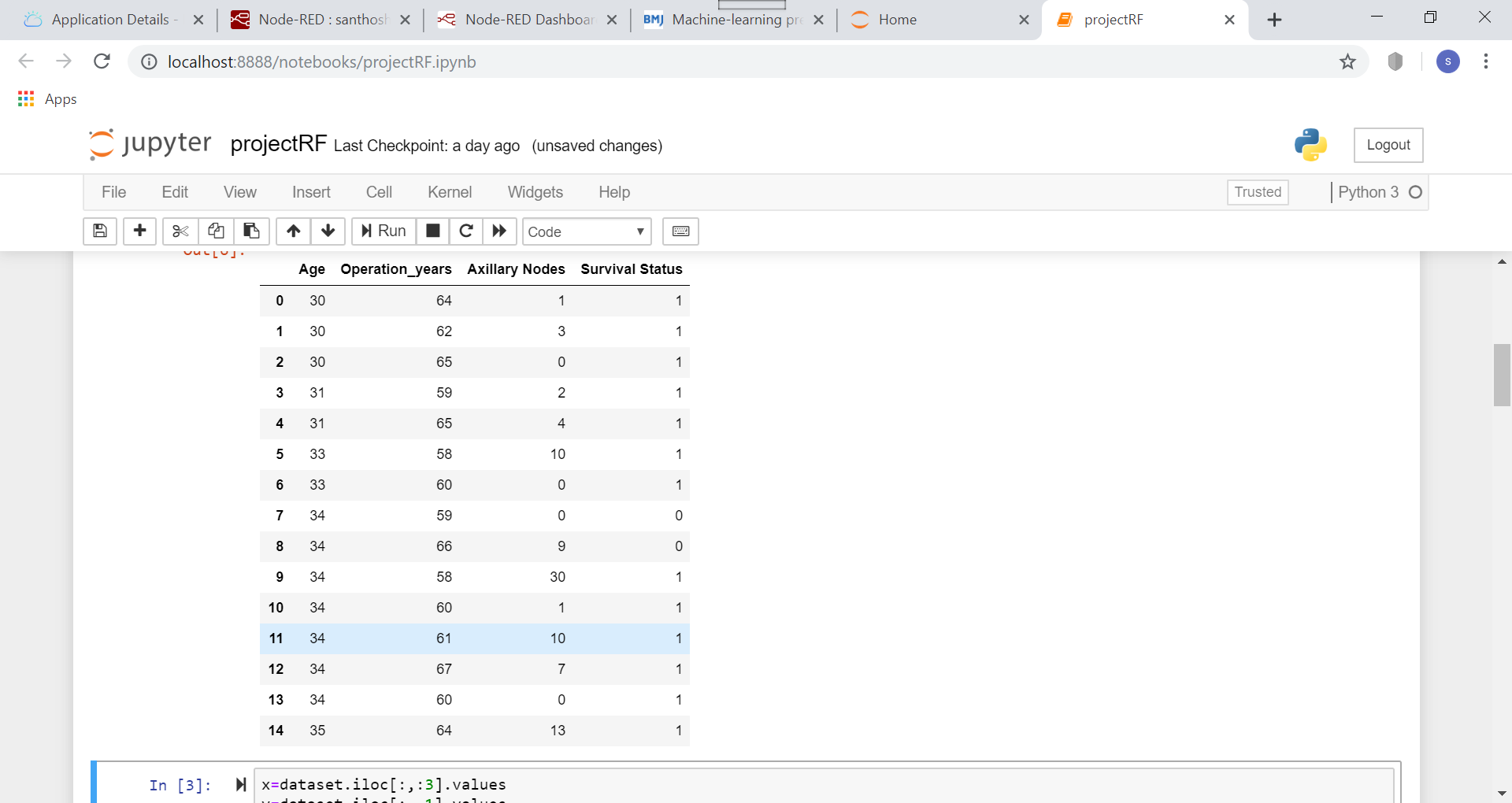
The Survival Analysis Cancer Prediction is used in Medical Organizations, Research Labs etc.

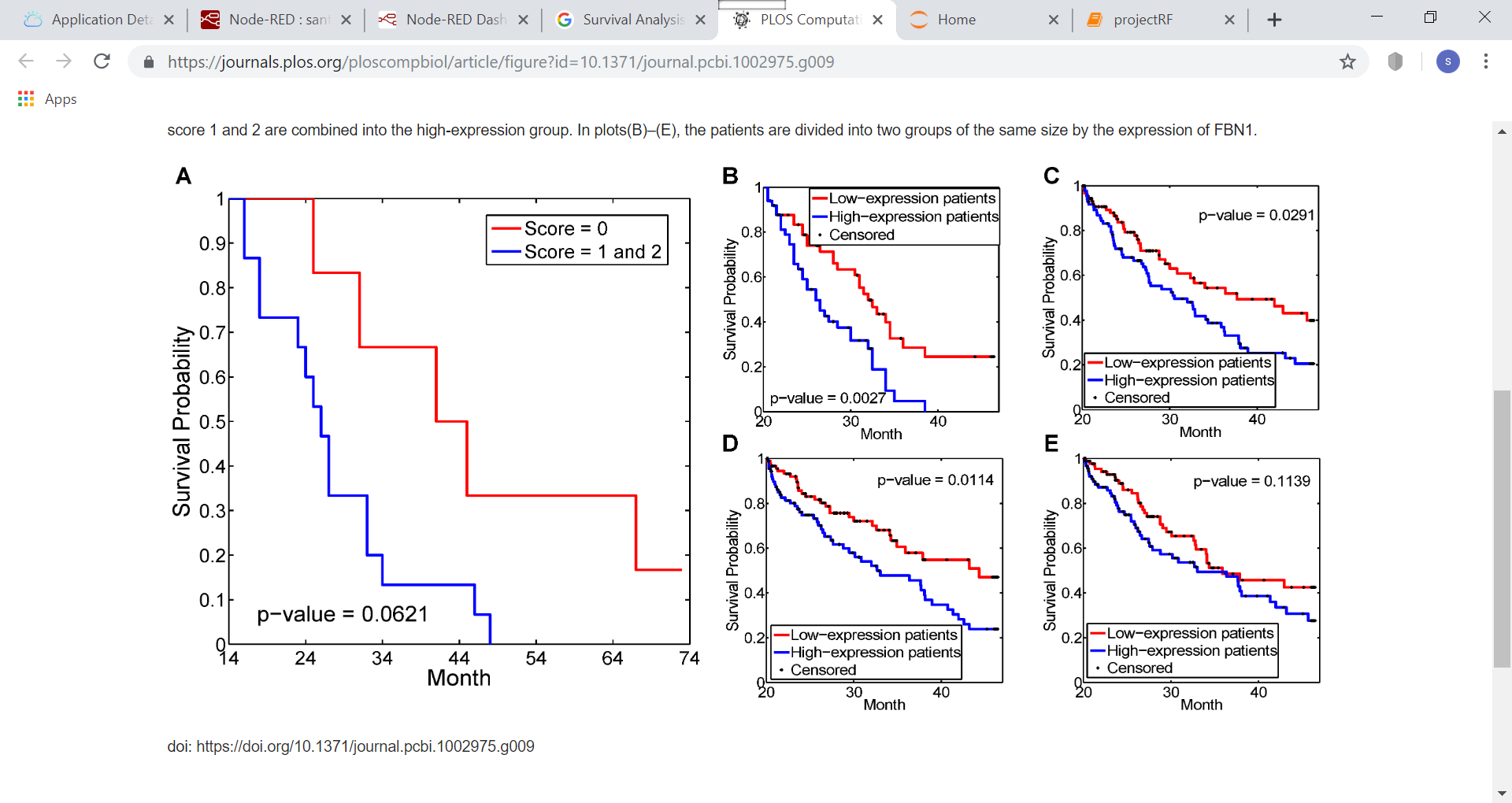
**Data Collection**

The Dataset of Survival Analysis Cancer Prediction is taken from <https://data.world.com> which has factors like age, operation\_years, axillary nodes, survival\_status of 300 persons. The input attributes used in this data are age, operation\_years, axillary nodes. Based on the above attributes we get an output attributes as survival\_status. In this model the X values are trained and then tested, based on these X\_test values we can predict Y values. Based on the slicing values of X&Y attributes we can test the X values. By using X\_test values along with “classifier” variable which is imported by sklearn package can predict the Y values The “classifier” variable is created based on X train and Y train values.

**Methodology**

**Figures and tables**

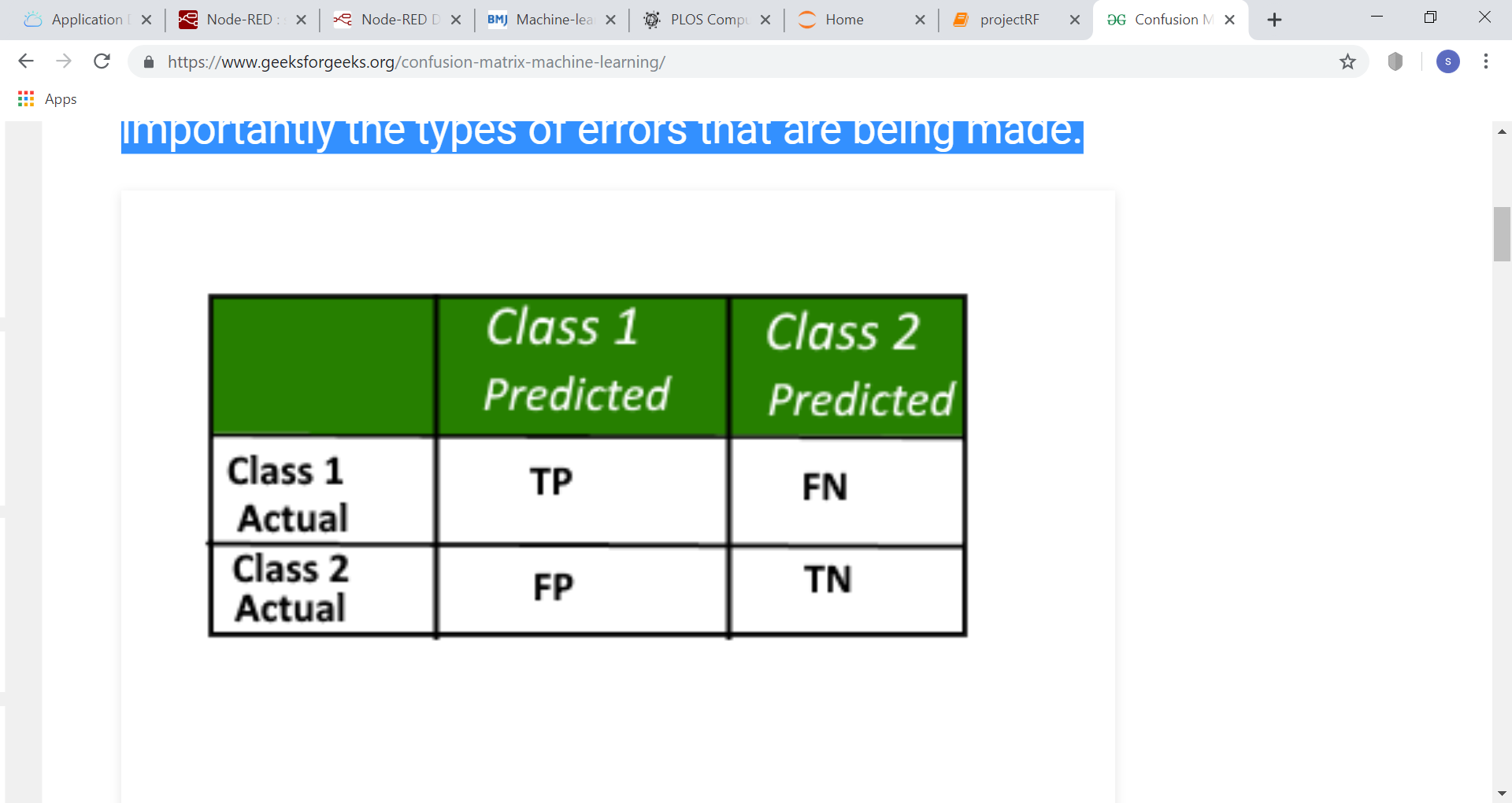
****

****

**Statistical techniques**

In the process first implemented step is: Splitting the dataset into the Training set and Test set from sklearn. model selection package.

* Test set in machine learning is a secondary (or tertiary) data set that is used to test a machine learning program after it has been trained on an initial training data set. The idea is that predictive models always have some sort of unknown capacity that needs to be tested out, as opposed to analyzed from a programming perspective A test set is also known as a test data set or test data.
* The idea of using training data in machine learning programs is a simple concept, but it is also very foundational to the way that these technologies work. The training data is an initial set of data used to help a program understand how to apply technologies like neural networks to learn and produce sophisticated results. It may be complemented by subsequent sets of data called validation and testing sets. Training data is also known as a training set, training dataset or learning set.
* Confusion matrix is used to display confusion matrices values, which has actual and predicted classes. If True positive and true negative values are more then accuracy and predicted values are proportional to it.
* A confusion matrix is a summary of prediction results on a classification problem.  
  The number of correct and incorrect predictions are summarized with count values and broken down by each class. This is the key to the confusion matrix.  
  The confusion matrix shows the ways in which your classification model is confused when it makes predictions. It gives us insight not only into the errors being made by a classifier but more importantly the types of errors that are being made.



Here,  
Class1:Positive  
 Class 2 : Negative

**Definition of the Terms:**

* Positive(P): Observation is positive (for example: is an apple).
* Negative(N): Observation is not positive (for example: is not an apple).
* True Positive(TP): Observation is positive, and is predicted to be positive.
* False Negative(FN): Observation is positive, but is predicted negative.
* True Negative(TN): Observation is negative, and is predicted to be negative.
* False Positive (FP): Observation is negative, but is predicted positive.

**Classification Rate/Accuracy:**Classification Rate or Accuracy is given by the relation  
****

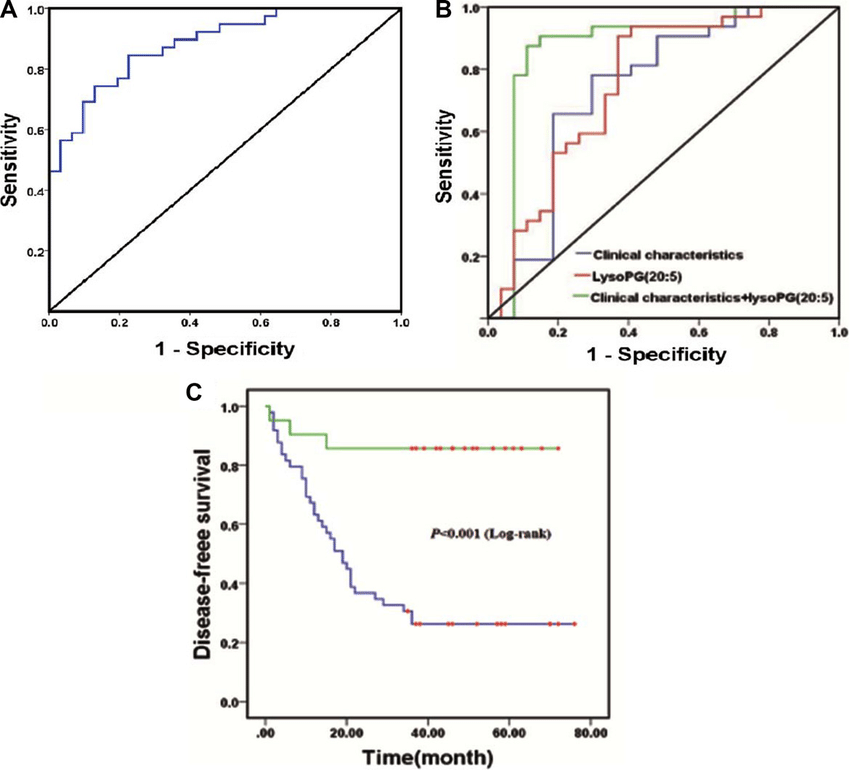
The sklearn. metrics package imports accuracy score method, accuracy\_score (), which returns “accuracy classification score”. What it does is the calculation of “How accurate the classification is.”

**Data Modelling**

For the survival analysis cancer prediction dataset, we used “Random forest classification” model. In this data to predict cancer there are only two values are available in output variable so this classification model is used. Random forest model is correctly suitable for this data than any other classification model where regression has no use to implement.

**Visualization**

This model mainly depends upon “ROC-AUC” curves. To construct the ROC graph we take fpr and tpr values which are present in confusion matrix. To plot an AUC curve threshold attribute must be defined.

****

**Findings and suggestions:**

A project like this has already been done, we have made changes and improvements to it and made it a better project. The improvements we have applied is random forest classification model which gives correct prediction and accuracy compared with other classification models.

**Review of Literature**

The data set for Survival analysis-cancer prediction is taken from

* <https://data.world.com>
* [https://data.gov.in](https://data.gov.in/)
* Huang J. 1999 Asymptotic properties of nonparametric estimation based on partly interval-censored data Statistica SinicaSubramanian.
* J, Simon RM. Gene expression-based prognostic signatures in lung cancer ready for clinical use? J Natl Cancer Inst. 2010.

**Conclusion**

Machine learning applied to information from a disease-specific (cancer) database and the EAR can be used to predict outcomes. Improved prediction of outcome has the potential to help clinicians make more meaningful decisions about treatment and to assist with planning of future social and care needs. Most importantly, the approach described makes use of digital data that is already routinely collected but underexploited by clinical health systems.