# Prediction of lung cancer patient survival via supervised machine learning classification techniques

Abhishek Vyas - MT19086
Chirag Chawla - MT19089

**Abstract : For lung cancer, the survival months are an important aspect to improve healthcare as well as it can be used to evaluate the patient prognosis.So in this project we are trying to predict the survival months for patients. For this we have used a lung cancer dataset from SEER and we have applied five well known machine learning algorithms in order to predict the survival months. We have used SVM, Random Forest, Ensemble regressor, Linear Regression, Gradient Boosting Machine. The best performing technique was the ensemble with a Root Mean Square Error (RMSE) value of 14.64.**

## I. Problem Background

Machine learning uses mathematical algorithms in order to extract the pattern from the large amount of dataset, and these patterns can be used to predict the outcomes for the unseen data. Nowadays machine learning is being used in all the fields which includes Finance, Insurance, Social media, fraud detection. But it is difficult to apply the machine learning algorithms on the biological dataset as these datasets are not available publicly. One exception is the Surveillance, Epidemiology, and End Results (SEER) program from the National Cancer Institute (NCI) at the National Institutes of Health (NIH). As the largest publicly available cancer dataset, this database provides de-identified information on cancer statistics of the United States population, thus facilitating large-scale outcome analysis. We can apply machine learning algorithms on these datasets.\\

This paper is basically trying to predict the survival time of the cancer patients. The goal of evaluating survivability is improving health care and providing the information to the patients and clinicians, as the survival time is important in order to evaluate the patient prognosis. We are using the lung cancer dataset from SEER which consists of emographic (e.g., age), diagnostic (e.g., tumor size), and procedural information (e.g., Radiation and/or Surgery applied) to fit our machine learning model.

## II. Dataset

The dataset is taken from the SEER which consists of only the records for lung cancer along with the diagnosis year lies between 2004 - 2009 (inclusive). The dataset consists of 18 attributes which includes emographic (age), diagnostic (tumor size), and procedural attributes (Radiation and/or Surgery applied).

## III. Preprocessing

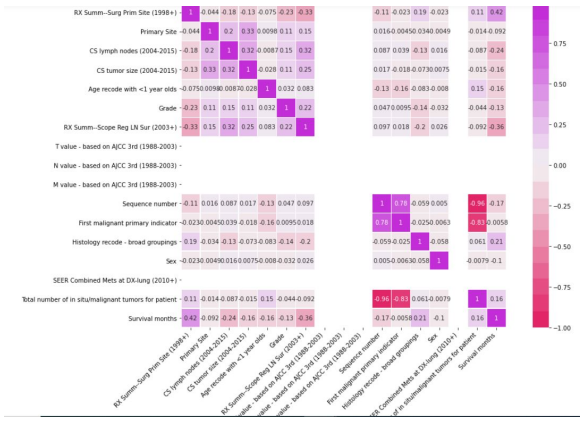
### *Label Encoder*

First of all various attributes (Age recode with <1 year olds, Grade, RX Summ--Scope Reg LN Sur (2003+), T value - based on AJCC 3rd (1988-2003), N value - based on AJCC 3rd (1988-2003), M value - based on AJCC 3rd (1988-2003), Sequence number, First malignant primary indicator, Histology recode - broad groupings, Sex, SEER Combined Mets at DX-lung (2010+), Total number of in situ/malignant tumors for patient) which are of type 'object', but machine learning algorithms work efficiently on numerical kind of attributes. So we have applied labelEncoder to convert these object type attributes to int type attributes.

### *Duplicates*

Then we have deleted the duplicate records from the dataset.

***Heatmap :***



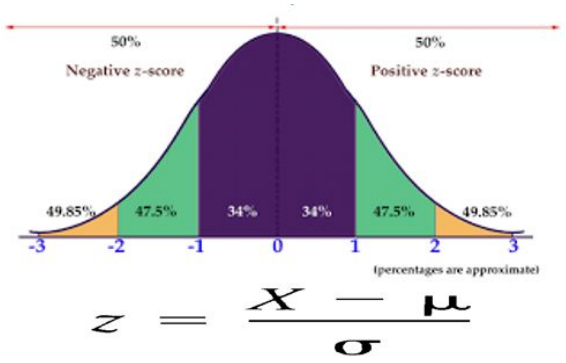
Heat map is created which represents the correlation of each attribute with every other attribute.

As we can see, 3 columns ( T-value, N-value, M-value) doesn't have any correlation with any other attributes.

So these 3 columns are deleted.

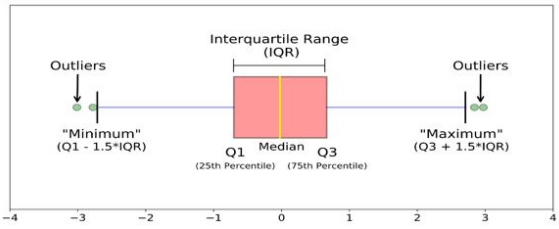
***Removing Outliers by Z-score & IQR :***

*Z - score :* We can find out the number of standard deviation values away from mean. Here in our case we took the data which is in range between -3sd to +3sd.

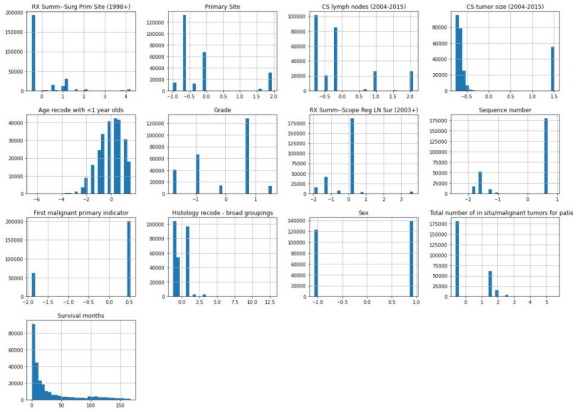



$$z = \frac{X - \mu}{\sigma}$$

*IQR (Interquartile Range) :* We detect the outliers using IQR which tells us the variation in the dataset. Here any value which is beyond the range for Q1-1.5*IQR to Q3+1.5*IQR is treated as an outlier where IQR = Q3 - Q1.




***Histogram Plots :***



## IV. Methods

Our task is to predict the survival months for the lung cancer patients. For this we have used 5 models :- Random Forest, SVM, Gradient Boosting Machine, Linear Regression, Custom Ensemble. And after predicting the survival time, the evaluation metric used is RMSE (root mean squared error) which is the root of average squared difference between the actual and predicted value, Standard deviation, standard deviation for residuals, mean squared error. And for validation, we are using 10 fold cross validation and in each validation iteration, we are training our model using the 70 percent of the data and 30 percent held out dataset is used to evaluate the RMSE score. Then we will take the mean of RMSE of these 10 folds.

***Random Forest :***

Random forest is an ensemble machine learning technique which uses the decision tree as its baseline model. So in random forest, it will try to construct multiple distinct individual decision trees (specified by n-estimator) from the training dataset. Then for the test data, we are using these trees to predict the values of test data. So at last we will get an "n_estimator" number of values for each test sample. Now from this, either we can classify by taking the

majority class or we can predict the continuous value by taking the mean of these values. In our project, we are using predefined random forest as a regressor from sklearn library with n_estimator=200 and max-depth=10.

***SVM :***

SVM (Support vector machine) is a non-probabilistic machine learning algorithm. For two class classification problems, SVM tries to draw a linear hyperplane which separates these two classes, and if the decision boundary is not linear it will introduce the kernel methods to extract the non-linear relation. Now this SVM can be used as the regression to predict the continuous value by taking the difference of test data points from the hyperplane.

***Gradient Boosting :***

Boosting is an ensemble modeling technique which attempts to build a strong classifier from the number of weak classifiers. It is done building a model by using weak models in series. Firstly, a model is built from the training data. Then the second model is built which tries to correct the errors present in the first model. This procedure is continued and models are added until either the complete training data set is predicted correctly or the maximum number of models are added.

***Decision Tree :***

As the name suggests, it is a tree kind of structure. Decision trees can be exploited as both classifiers as well as regressors. A node in the decision tree represents some attribute of the training dataset. And the leaf node represents a certain class label. Given a dataset the optimized decision tree is constructed through various algorithms. From a given training sample and the decision tree, it starts from root and based on the value of root attribute of the test sample it will branch to one of the children of root and do the same procedure till it reaches the leaf node which corresponds to a certain class label. We are using the decision tree Regressor from sklearn Library.

***Linear Regression :***

Linear regression is a simple and mostly used machine learning algorithm which is used for the regression problem. Linear regression basically tries to extract the linear relation between the dependent variables and the independent variable by learning their coefficients and the bias. Linear regression is of two types :

Univariate linear regression which is having one dependent variable (X) and one independent variable and linear regression tries to optimize the linear line coefficients (W1, W0).

Multivariate Linear Regression, which is having 'n' independent variables and one dependent variable and it tries to optimize the coefficients (Wn ,...., W1, W0).

Now based on these coefficients, we will try to predict the value for test data samples by : Xn*Wn + .... + X1*W1 + W0. In our project we are using multivariate linear regression from sklearn library.

***Ensemble :***

Ensemble is a technique to ensemble (combine) multiple machine learning algorithms to predict the desired output. If you apply only a single model, then there is a chance of overfitting of the model towards the train dataset. Despite overfitting, there might be the chances of biases i.e model can be biased towards the training dataset. So in order to remove these problems, ensemble classifiers or regressors are being used. So the basic concept of ensemble regressor is that we are not relying on a single model to predict the outcomes, instead ensemble regressors are fitting multiple models on a train dataset, and predict the test data sample outcome based on these multiple models.

## V. Results and Analysis

We have applied various machine learning techniques and evaluated various metrics on the predicted value such as RMSE (Root mean squared error), Standard Deviation, Standard deviation of residuals, Mean absolute Error (MAE), Mean squared error (MSE). So the results are mentioned below :

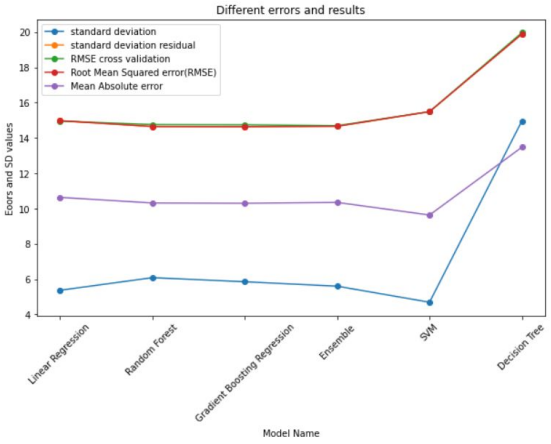
| ML Algorithm | Standard Deviation | Standard Deviation of Residuals | RMSE | MAE | MSE |
|---|---|---|---|---|---|
| Linear Regression | 5.371178 | 14.978768 | 14.978102 | 10.634822 | 224.343545 |
| Random Forest | 6.086403 | 14.649574 | 14.648924 | 10.316861 | 214.590961 |
| Gradient Boosting | 5.855198 | 14.641853 | 14.641202 | 10.300802 | 214.364804 |
| Ensemble | 5.603182 | 14.662370 | 14.661718 | 10.349254 | 214.965979 |
| SVM | 4.694388 | 15.490977 | 15.490289 | 9.634750 | 239.949047 |
| Decision Tree | 14.965710 | 19.888878 | 19.887995 | 13.477005 | 395.532334 |

Also we have analysed these results using the plot shown below, here are some of the analysis:

- RMSE and standard deviation of residuals are approximately the same ranging from

14-20 for various machine learning algorithms.

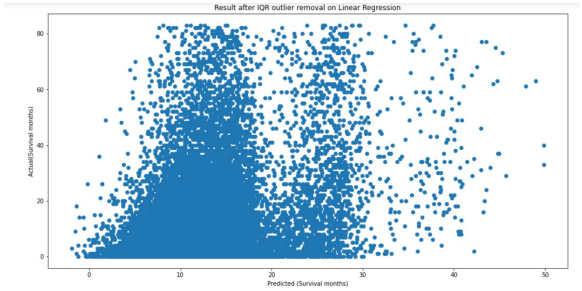
- Mean Absolute Error is ranging from 10-14 for various machine learning algorithms.
- Standard Deviation is ranging from 4-14 for various machine learning algorithms.



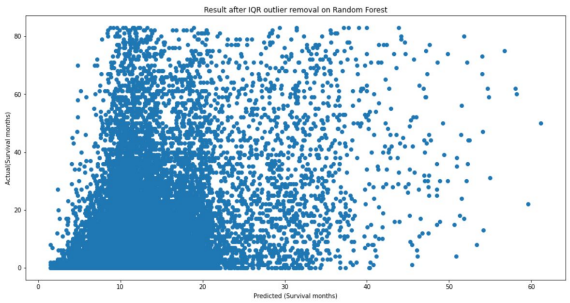

## VI. Comparison of predicted to actual survival time

We have plotted the scatter plot in order to compare the predicted value with the actual value for all the six models that we have implemented.

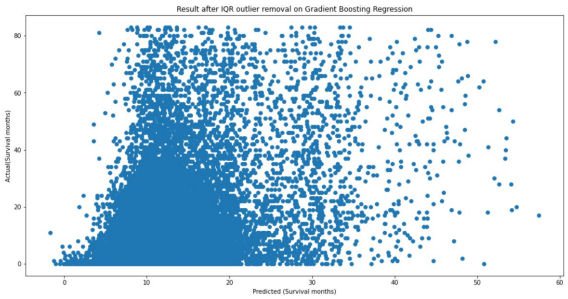
### *Linear Regression :*



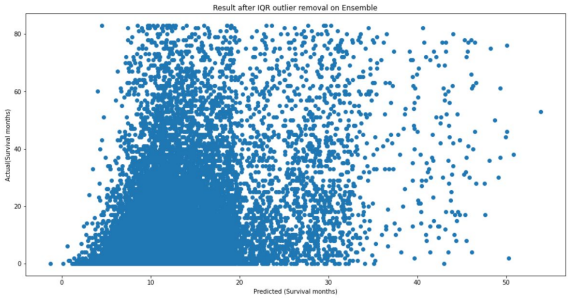

### *Random Forest :*



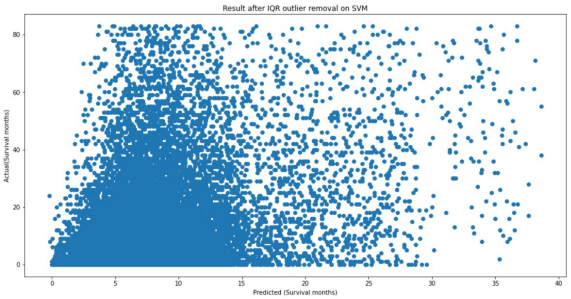

### *Gradient Boosting :*



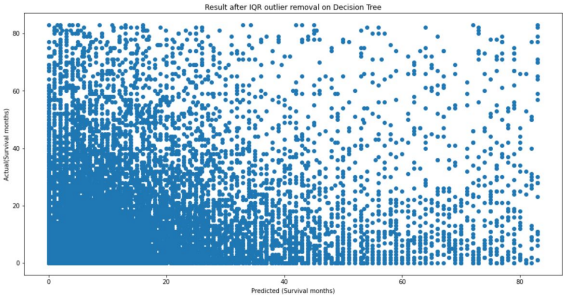

### *Ensemble Technique :*




### *Support Vector Regressor (SVR) :*




### *Decision Tree:*



## VII. Conclusion

We have implemented six different machine learning models in order to predict the Survival month for patients. We are able to achieve the best performance using the Gradient Boosting Technique with RMSE 14.64 which is also the improvement of that of the RMSE of the paper which is 15.32.

## VIII. References


[1] NCI_SEER_Training_Lung_Cancer_Stats, Introduction to Lung Cancer. SEER Training Modules, National Cancer Institute, 2015 (Available from: http://training.seer. cancer.gov/lung/ ).

[2] NCI_SEER_Overview, Overview of the SEER Program. Surveillance Epidemiology and End Results, (2015) (Available from: http://seer.cancer.gov/about/ ).

[3] SEER_Program. Surveillance, Epidemiology, and End Results (SEER) Program (www.seer.cancer.gov) Research Data 1973–2009, National Cancer Institute, DCCPS, Surveillance Research Program, Surveillance Systems Branch, released April 2012, based on the November 2011 submission.

[4] ACS_Cancer_Facts, Cancer Facts & Figures, American Cancer Society, 2015.

[5] NCI_Lung_Cancer_Info, What You Need To Know About Lung Cancer, National Cancer Institute, 2015 (Available from: http://www.cancer.gov/publications/ patient-education/wyntk-lung-cancer ).

[6] NCI_Lung_Cancer_Overview, Lung Cancer, National Cancer Institute, 2015 (Available from: http://www.cancer.gov/cancertopics/types/lung/ ).

[7] C. Clément-Duchêne, C. Carnin, F. Guillemin, Y. Martineta, How accurate are physicians in the prediction of patient survival in advanced lung cancer? Oncologist 15 (2010) 782–789.

[8] M.F. Muers, P. Shevlin, J. Brown, Prognosis in lung cancer: physicians' opinions compared with outcome and a predictive model, Thorax 51 (1996) 894–902.

[9] S. Ramalingam, K. Pawlish, S. Gadgeel, R. Demers, G. Kalemkerian, Lung cancer in young patients: analysis of a Surveillance, Epidemiology, and End Results database, J. Clin. Oncol. 16 (1998) 651–657.

[10] T.K. Owonikoko, C.C. Ragin, C.P. Belani, A.B. Oton, W.E. Gooding, E. Taioli, et al., Lung cancer in elderly patients: an analysis of the surveillance, epidemiology, and end results database, J. Clin. Oncol. 25 (2007) 5570–5577.

[11] A. Bhaskarla, P.C. Tang, T. Mashtare, C.E. Nwogu, T.L. Demmy, A.A. Adjei, et al., Analysis of second primary lung cancers in the SEER database, J. Surg. Res. 162 (2010) 1–6.

[12] M.J. Hayat, N. Howlader, M.E. Reichman, B.K. Edwards, Cancer statistics, trends, and multiple primary cancer analyses from the Surveillance, Epidemiology, and End Results (SEER) Program, Oncologist 12 (2007) 20–37.

[13] M.J. Thun, L.M. Hannan, L.L. Adams-Campbell, P. Boffetta, J.E. Buring, D. Feskanich, et al., Lung cancer occurrence in never-smokers: an analysis of 13 cohorts and 22 cancer registry studies, PLoS Med. 5 (2008) e185.

[14] J.B. Fu, T.Y. Kau, R.K. Severson, G.P. Kalemkerian, Lung cancer in women: analysis of the national surveillance, epidemiology, and end results database, CHEST J. 127 (2005) 768–777.

[15] X. Wu, V. Chen, J. Martin, S. Roffers, F. Groves, C. Correa, et al., Comparative Analysis of Incidence Rates Subcommittee, Data Evaluation and Publication Committee, North American Association of Central Cancer Registries. Subsite-specific colorectal cancer incidence rates and stage distributions among Asians and Pacific Islanders in the United States, 1995–1999, Cancer Epidemiol. Biomarkers Prev. 13 (2004) 1215–1222.

[16] S.J. Wang, C.D. Fuller, R. Emery, Thomas Jr CR: Conditional survival in rectal cancer: a SEER database analysis, Gastrointest. Cancer Res.: GCR 1 (2007) 84.

[17] Identifying hotspots in lung cancer data using association rule mining, in:

A. Agrawal, A. Choudhary (Eds.), 11th International Conference on Data Mining
Workshops (ICDMW), IEEE, 2011.
[18] Finding survival groups in SEER lung cancer data. machine learning and applications
(ICMLA), in: I. Skrypnyk (Ed.), 11th International Conference On; 2012, IEEE,
2012.
[19] A lung cancer outcome calculator using ensemble data mining on SEER data, in:
A. Agrawal, S. Misra, R. Narayanan, L. Polepeddi, A. Choudhary (Eds.), Proceedings
of the Tenth International Workshop on Data Mining in Bioinformatics, ACM, 2011.
[20] A. Agrawal, A. Choudhary, Association rule mining based HotSpot analysis on SEER
lung cancer data, Int. J. Knowl. Discov. Bioinf. (IJKDB) 2 (2011) 34–54.
[21] N. Kapadia, F. Vigneau, W. Quarshie, A. Schwartz, F. Kong, Patterns of practice and
outcomes for stage I non-small cell lung cancer (NSCLC): analysis of SEER-17 data,
1999–2008, Int. J. Radiat. Oncol.* Biol.* Phys. 84 (2012) S545.
[22] A clustering-based approach to predict outcome in cancer patients, in: K. Xing,
D. Chen, D. Henson, L. Sheng (Eds.), Sixth International Conference on Machine
Learning and Applications (ICMLA), IEEE, 2007.
[23] T. Hastie, R. Tibshirani, J. Friedman, The Elements of Statistical Learning, Springer-
Verlag, New York, 2001.
[24] T.S. Madhulatha, An overview on clustering methods, IOSR J. Eng. 2 (2012)
719–725.
[25] D. Delen, G. Walker, A. Kadam, Predicting breast cancer survivability: a comparison
of three data mining methods, Artif. Intell. Med. 34 (2005) 113–127.
[26] D. Delen, Analysis of cancer data: a data mining approach, Expert Syst. 26 (2009)
100–112.
[27] Knowledge extraction from prostate cancer data, in: D. Delen, N. Patil (Eds.), 39th
Hawaii International Conference on System Sciences, Hawaii, 2006.
[28] N.A. Noohi, M. Ahmadzadeh, M. Fardaer, Medical data mining and predictive
model for colon cancer survivability, Int. J. Innov. Res. Eng. Sci. (2013) 2.
[29] Colon cancer survival prediction using ensemble data mining on SEER data, in:
R. Al-Bahrani, A. Agrawal, A. Choudhary (Eds.), IEEE Big Data Workshop on
Bioinformatics and Health Informatics (2013).
[30] D. Chen, K. Xing, D. Henson, L. Sheng, A.M. Schwartz, X. Cheng, Developing
prognostic systems of cancer patients by ensemble clustering, J. Biomed.
Biotechnol. 2009 (2009) (632786).
[31] D. Fradkin, Machine Learning Methods in the Analysis of Lung Cancer Survival
Data, (2006) (February).
[32] C.M. Lynch, V.H. van Berkel, H.B. Frieboes, Application of unsupervised analysis
techniques to lung cancer patient data, PLoS One 12 (2017) e0184370.
[33] V. Krishnaiah, G. Narsimha, N.S. Chandra, Diagnosis of lung cancer prediction
system using data mining classification techniques, Int. J. Comput. Sci. Inf. Technol.
4 (2013) 39–45.
[34] G. Dimitoglu, J.A. Adams, C.M. Ji, Comparison of the C4: 5 and a naive bayes
classifier for the prediction of lung cancer survivability, J. Comput. 4 (2012) 1–9.
[35] A. Agrawal, S. Misra, R. Narayanan, L. Polepeddi, A. Choudhary, Lung cancer
survival prediction using ensemble data mining on SEER data, Sci. Program. 20
(2012) 29–42.
[36] Mining association rules between sets of items in large databases, in: R. Agrawal,
T. Imieliński, A. Swami (Eds.), SIGMOD Record, ACM, 1993.
[37] Y. Wu, Propensity Score Analysis to Compare Effects of Radiation and Surgery on
Survival Time of Lung Cancer Patients from National Cancer Registry (SEER)
[Master's], Epidemiology and Biostatistics: School of Public Health, SUNY-Albany,
2006.
[38] R Documentation Control for Rpart Fits. Package Rpart Version 41-10, (2017).