Machine Learning for Survival Analysis: A New Approach

David Dooling[1,☯], Angela Kim[1,‡], Jennifer Webster[1,☯]

1 Innovative Oncology Business Solutions, Albuquerque, NM, USA

☯These authors contributed equally to this work.
‡These authors also contributed equally to this work.
*** ddooling@innovativeobs.com**

Abstract

We have applied a little-known data transformation to subsets of the Surveillance, Epidemiology, and End Results (SEER) publically available data of the National Cancer Institute (NCI) to make it suitable input to standard machine learning classifiers. This transformation properly treats the right-censored data in the SEER data and the resulting Random Forest and Multi-Layer Perceptron models predict full survival curves. Treating the 6, 12, and 60 months points of the resulting survival curves as 3 binary classifiers, the 18 resulting classifiers have AUC values ranging from .765 to .885. Further evidence that the models have generalized well from the training data is provided by the extremely high levels of agreement between the random forest and neural network models predictions on the 6, 12, and 60 month binary classifiers.

Author Summary

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Curabitur eget porta erat. Morbi consectetur est vel gravida pretium. Suspendisse ut dui eu ante cursus gravida non sed sem. Nullam sapien tellus, commodo id velit id, eleifend volutpat quam. Phasellus mauris velit, dapibus finibus elementum vel, pulvinar non tellus. Nunc pellentesque pretium diam, quis maximus dolor faucibus id. Nunc convallis sodales ante, ut ullamcorper est egestas vitae. Nam sit amet enim ultrices, ultrices elit pulvinar, volutpat risus.

Introduction

Opportunities are emerging in many indutries today to develop and deploy services that cater to individual needs and preferences. Music afficianados can create their own radio stations tailored to their individual tastes from Pandora[1], bibliophiles can receive highly trustworthy book recommendations from goodreads.com[2], and Google will provide directions between any two points, giving options such as mode of transportation and as well as warnings of delays in realtime.[3] These individualized services share many

[1]Pandora Internet Radio - Listen to Free Music You'll Love, `http://www.pandora.com/` (accessed 27 Jan 2016)

[2]Share Book Recommendations With Your Friends, Join Book Clubs, Answer Trivia, `https://www.goodreads.com/` (accessed 27 Jan 2016)

[3]Google Maps, `https://goo.gl/lD7Jwf` (accessed 27 Jan 2016)

common features. In particular, they leverage large databases of aggregated information to learn and extract information relevant to individuals. Extracting actionable information from data is changing the fabric of modern business. A class of techniques that transforms data into actionable information goes by the name of Machine Learning [1]. Machine Learning has recently become a popular method to answer questions and solve problems that are too complex to solve via traditional methods.

The primary objective of this study is to show how machine learning methods can be trained with data in cancer registries to produce personalized survival prognosis curves, but the methods presented below can be applied to any type of survival data. Traditionally, cancer survival curves have been estimated using Kaplan-Meier methods [2]. Kaplan-Meier methodology also uses large datasets to make predictions, but the resulting information is not personal; the resulting curves are summaries for a population and not necessarily relevant or particularly accurate for any given individual. This propery of Kaplan-Meier methods is exacerbated when dealing with heterogeneous populations. The methods described below also take full advantage of all relevant aggregate information, but are able to provide personalized survival curves relevant to individual subjects. This objective is in keeping with the recent movement in medicine known as Predictive, Preventive and Personalized Medicine (PPPM), which aims to leverage increasing amounts of health related data to maximize quality of care and to intelligenctly eliminate inefficient and unecessary use of resources [3]. This capability of providing individualized survival curve prognosis is a direct result of the recent advances in computing power and machine learning algorithms, and similar methodology is becoming commonplace in many industries. These techniques are now infiltrating the healthcare industry, in spite of some of the data aggregation challenges posed by the Health Insurance Portability and Accountability Act (HIPPA) of 1996. This study makes use of a freely available data source that circumvents the restrictions imposed by HIPPA.

The Surveillance, Epidemiolgy, and End Results (SEER) Program of the National Cancer Institute (NCI) has been collecting data because intuitively researchers feel confident that this data will eventually allow researches to detect information crucial to patients and providers including the relationships between the types of data collected (demographic as well as staging information, treatment and disease characteristics) and the survival outcomes. Though these relationships evade capture by traditional methods, it is possible to surface them with two machine learning techniques known as *Random Forests* and *Neural Networks*. As will be demonstrated in section , these two methods produce very similar results when applied to the SEER dataset, and are based on almost diametrically opposed learning philosophies, which lends confidence in the validity of the results.

The Surveillance, Epidemiolgy, and End Results (SEER) Program of the National Cancer Institute (NCI) is the most recognized authoritative source of information on cancer incidence and survival in the United States. SEER currently collects and publishes cancer incidence and survival data from population-based cancer registries covering approximately 28 percent of the US population.

Quoting directly from the SEER website [4]:

> The SEER program registries routinely collect data on patient demographics, primary tumor site, tumor morphology and stage at diagnosis, first course of treatment, and follow-up for vital status. This program is the only comprehensive source of population-based information in the United States that includes stage of cancer at the time of diagnosis and patient survival data. The mortality data reported by SEER are provided by the National Center for Health Statistics. The population data used in calculating cancer rates is obtained periodically from the Census Bureau.

> Updated annually and provided as a public service in print and electronic formats, SEER data are used by thousands of researchers, clinicians, public health officials, legislators, policymakers, community groups, and the public.

One characterstic of the SEER data that is shared by many datasets in the medical field goes by the name of "censored data." Observations are labeled censored when the survival time information is incomplete. The SEER data contains the number of months each patient survived, as well as an indicator variable showing whether or not the patient is still alive at the end of the data collection period. Methods to deal effectively with this kind of "right-censored data" include Kaplan-Meier curves and Cox Proportional Hazard models [2]. The Kaplan-Meier techniques only give estimates for cohorts of patients and are not applicable for predicting the surival curve for a single patient, and the Cox Proportional Hazard models require a fairly restrictive set ot assumptions to be satisifed in order to yield reliable results.

Previous work applying machine learning methods to subsets of the SEER data include creative attempts to deal with the problems presented by "right-censored data." Shin et al. [5] use semi-supervised learning techniques to predict 5 year survival, essentially imputing values for SEER records where the survival months infomation is censored at a value less than 5 years. Zolbanin et al. [6] investigate the effects of comordbidities; i.e., patients with two different cancer diagnosises, but their treatment of the censored data underestimates the survival probabilities. All records representing patients who survived at least 60 months as well as all those who died earlier than 60 months were considered, but patients alive prior to 60 months but censored out of the study before 60 months were not included. This treatment biases the data and the predictions, leading to overly pessimistic survival probablilites predicted by the models.

Previous work applying machine learning methods based on decision trees to survival data in general have a long history, starting with Gordon et al. [7]. A summary of more recent developments concerning *survival trees* is provided by Bou-Hamad et al. [8]. These methods focus on altering the splitting critieria used in decision tree growth to account for the censoring, and use 1958 Kaplan-Meier methods at the resulting nodes for prediction purposes. These methods do not generalize to non-tree-based machine learning algorithms, though Ishwaran et al. have extended the methodology to *random survival forests*, ensembles of *survival trees* [9].

IOBS has applied a little-known technique to transform the SEER data to make it amenable to more powerful machine learning methods. Instead of modifying existing learning algorithms in drastic ways, we focus attention on the input data. This approach allows for different machine learning algorithms to use the same data with no modification. The essential idea is to recast the problem to an appropriate discrete classification problem instead of a regression problem (predicting survival months). Treating months after diagnosis as just another discrete feature, the SEER data (or any other right-censored data) can be transformed to make predictions for the hazard function (probability of dying in the next month, given that the patient has not yet died). The full survival function can then be derived from the hazard function.

This paper is organized as follows. We introduce the subsets of the SEER data used for this study, and present survival curves computed from traditional methods based on this data for the three cancer types *lung*, *breast*, and *colon*. We then present the essential methodology of this work, the data trasnformation that allows censored survival data to be used as input to exisiting machine learning classifiers. Then we present the details of the trained models, including some some subtleties arising from the data transformation pertaining to the partition into training and test datasets. The method of deriving binary classifiers from the models' predictions for the survival curves is presented. In this paper, we have constructed binary classifiers corresponding to 6, 12, and 60 months, as these are standard metrics in cancer survival prognosis. Then follows

a dicussion of the evaluation of the trained models. The performance metrics are the 18 AUC curves associated with the 6, 12, and 60 month survival binary classifiers for the two models associated with each cancer type. We also present additional evidence supporting validity of the predictions by computing the levels of agreement between the random forest and neural network models for each of the 18 binary classifiers and find striking agreement. Next we provide urls for 6 web applications that use the trained models to predict individual cancer survival prognosis curves. These apps are hosted on the popular Heroku website, and allow for exploration of the nonlinear relationships between the input features and resulting survival prognosis. It is exactly these kinds of tools that are the goal of Predictive, Preventitive and Personalized Medicine. Finally, we present avenues for future research.

Materials and Methods

For this study we use the publically available 1973-2012 SEER incidence data files corresponding to colon, breast and lung cancer contained in the list below. SEER requires that researchers submit a request for the data, which includes an agreement form. Detailed documentation explaining the contents of both the incidence data files used in this study as well as a data dictionary for the 1973-2012 SEER incidence data files are available without the need to register or submit a data request [10].

- incidence\yr1973_2012.seer9\COLRECT.txt
- incidence\yr1973_2012.seer9\BREAST.txt
- incidence\yr1973_2012.seer9\RESPIR.txt
- incidence\yr1992_2012.sj_la_rg_ak\COLRECT.txt
- incidence\yr1992_2012.sj_la_rg_ak\BREAST.txt
- incidence\yr1992_2012.sj_la_rg_ak\RESPIR.txt
- incidence\yr2000_2012.ca_ky_lo_nj_ga\COLRECT.txt
- incidence\yr2000_2012.ca_ky_lo_nj_ga\BREAST.txt
- incidence\yr2000_2012.ca_ky_lo_nj_ga\RESPIR.txt
- incidence\yr2005.lo_2nd_half\COLRECT.txt
- incidence\yr2005.lo_2nd_half\BREAST.txt
- incidence\yr2005.lo_2nd_half\RESPIR.txt

Data preparation and preprocessing

A great deal of data munging is necessary before using these SEER incidence files as input into machine learning algorithms. A preprocessing step common to each of the three cancer types studied involves the SEER `STATE-COUNTY RECODE` variable. The `STATE-COUNTY RECODE` field is a state-county combination where the first two characters represent the state FIPS code and the last three digits represent the FIPS county code. The FIPS code is a five-digit Federal Information Processing Standard (FIPS) code which uniquely identifies counties and county equivalents in the United States, certain U.S. possessions, and certain freely associated states. This particular field illustrates an important characteristic of machine learning, that is, the difference between *categorical features* and *numeric features*. All input into a machine learning algorithm must be numeric, but real numbers carry with them the usually extremely useful property known as the well-ordering property. Machine learning algorithms use the well-ordering property of the real numbers to learn. But if one is tasked with encoding a categorical feature into suitable numeric format for machine learning, it is necessary to do so in a way that removes the well-ordering property [11].

As a simple example of how to correctly treat categorical variables in a machine learning context, consider the SEER variable `SEX` . This variable is encoded in the

Code	Description
1	Male
2	Female

Table 1. Encoding of gender in the SEER incidence files. These types of categorical variables need to be transformed via one-hot-encoding.

SEER raw data files with a numeric 1 for males and a numeric 2 for females as shown in Table (1). Values such as "Male" and "Female" encoded as numbers are dangerous because if not handled properly, they can generate bogus results [12]. Leaving the infomation for `SEX` as in Table (1) implies that Female is somehow greater than Male. This implied ordering affects the machine learning algorithms' convergence on a model. Simply encoding Male by 2 and Female by 1 would result in a comletely different model, because of the now completely reversed ordering implied in the `SEX` variable. The proper way to transform the SEER `SEX` variable is to create two additional variables: `sex_Male` and `sex_Female`, and then to eliminate the variables `SEX` and `sex_Male` (keeping both of the variables `sex_Male` and `sex_Female` is a redundant represetation). For example,

Sex	→	sex_Male	sex_Female	→	sex_Female
1		1	0		0

(1)

and

Sex	→	sex_Male	sex_Female	→	sex_Female
2		0	1		1

(2)

The procedure outlined in Equations (1, 2) is known as one-hot encoding and needs to be applied to all of the nominal categorical variables in the SEER data that we wish to include in our predictive models. In particular, in order to include the geophgraphical information contained in the SEER categorical variable `STATE-COUNTY RECODE`, it becomes necessary to create a new feature variable for each of the distinct (state,county) pairs in the data. In the United States, there are approximately 3,000 counties. Clearly, transforming the `STATE-COUNTY RECODE` data representation into distinct (state_county) columns will explode the dataset to become wider than is optimal for machine learning. Adding extra columns to your dataset, making it wider, requires more data rows (making it taller) in order for machine learning algorithms to effectively learn [11]. Because one-hot coding `STATE-COUNTY RECODE` would cause such drastic shape changes in our data, we wish to avoid doing so. Fortunately, this variable, though given as a categorical variable, is actually a recode for three ordinal variables. There is an ordering among the (state_county) columns, namely longitude, latitude, and elevation. We can transform the data in `STATE-COUNTY RECODE` into three new numerical columns: `lat`, `lng`, and `elevation`.

For example, Table (2) shows how five entries of `STATE-COUNTY RECODE` corresponding to counties within New Mexico can be represented by the `elevation`, `lat`, and `lng` features.

It is a simple exercise to construct the full lookup table from the SEER `STATE-COUNTY RECODE` variable to the corresponding three values `elevation`, `lat`, and `lng`. We use the publically available dafafile from the United States Census Bureau [13] to map the state FIPS and county FIPS codes to query strings like those in the `address` field in Table (2). It is then possible to programmatically query the

Table 2. Example of the transformation of `STATE-COUNTY RECODE` to `elevation` , `lat` , and `lng` .

STATE-COUNTY RECODE	address	elevation	lat	lng
35001	Bernalillo+county+NM	5207.579772	35.017785	-106.629130
35003	Catron+county+NM	8089.242628	34.151517	-108.427605
35005	Chaves+county+NM	3559.931671	33.475739	-104.472330
35006	Cibola+county+NM	6443.415570	35.094756	-107.858387
35007	Colfax+county+NM	6147.749089	36.579976	-104.472330

Google Maps Geocoding API for the latitude and longitude [14], and the Google Maps Elevation API for the corresponding elevation [15]. An added benefit of this shift from the single categorical variable `STATE-COUNTY RECODE` to the three continuous numerical variables `lat` , `lng` , and `elevation` is that input into the web applications described later are not restricted to the states and counties coverered in the SEER registries; in fact, the input to the models can be any address you would enter into Google Maps and calls to the Google Maps Geocoding API and the Google Maps Elevation API provide the conversion from the address string to the input variables `lat` , `lng` , and `elevation` . The full lookup table analogous to Table (2) is available from a GitHub repository containing supplemental information for this study [16].

This study focused on three different cancer types, namely colorectal cancer, lung cancer, and breast cancer. In the SEER data, there are instances of subjects with multiple rows; whenever a subject, or patient, is diagnosed with a new tumor, an additional record is added. In this study, we restrict attention to the data corresponding to the first record of each subject; i.e., we wish to make models that predict survival prognosis based on the data available right after diagnosis. The full set of conditions defining the subsets of the SEER data used in this study follows below.

The four COLRECT.txt files were imported into a pandas DataFrame object. This data was then filtered according to the conditions in Table (3). The RESPIR.txt and BREAST.txt files were imported into separate dataframes in similar fashion and filtered according to the conditions in Table (4) and Table (**??**), respectively. The SEER variable `CS TUMOR SIZE` records the tumor size in millimeters if known. But if not known, `CS TUMOR SIZE` is given as '999', to indicate that the tumor size is "Unknown; size not stated; not stated in pateint record." In this study, we discard those records, as indicated in Tables (**??**, 3, 4).

Table 3. Filters applied to the Colon Cancer data.

Column	Filter
`SEQUENCE NUMBER-CENTRAL`	$\neq$ `"Unspecified"`
`AGE AT DIAGNOSIS`	$\neq$ `"Unknown age"`
`BIRTHDATE-YEAR`	$\neq$ `"Unknown year of birth"`
`YEAR OF DIAGNOSIS`	≥ 2004
`SURVIVAL MONTHS FLAG`	= `"1"`
`CS TUMOR SIZE EXT/EVAL`	$\neq$ `""`
`CS TUMOR SIZE`	$\neq 999$
`SEER RECORD NUMBER`	$= 1$
`PRIMARY SITE`	= `"LARGE INTESTINE, (EXCL. APPENDIX)"`
`SEQUENCE NUMBER-CENTRAL`	$= 0$

Table 4. Filters applied to the Lung Cancer data.

Column	Filter
SEQUENCE NUMBER-CENTRAL	$\neq$ "Unspecified"
AGE AT DIAGNOSIS	$\neq$ "Unknown age"
BIRTHDATE-YEAR	$\neq$ "Unknown year of birth"
YEAR OF DIAGNOSIS	≥ 2004
SURVIVAL MONTHS FLAG	= "1"
CS TUMOR SIZE EXT/EVAL	$\neq$ ""
CS TUMOR SIZE	$\neq 999$
SEER RECORD NUMBER	$= 1$
PRIMARY SITE	= "LUNG & BRONCHUS"
SEQUENCE NUMBER-CENTRAL	$= 0$

The following categorical features were one-hot encoded for each of the three datasets:

- SEX ,
- MARITAL STATUS AT DX ,
- RACE/ETHNICITY ,
- SPANISH/HISPANIC ORIGIN ,
- GRADE ,
- PRIMARY SITE ,
- LATERALITY ,
- SEER HISTORIC STAGE A ,
- HISTOLOGY RECODE--BROAD GROUPINGS ,
- MONTH OF DIAGNOSIS ,
- VITAL STATUS RECODE ,

and the STATE-COUNTY RECODE variable was dropped and replaced with the elevation , lat , and lng variables for all three datasets as illustrated in Table (2).

Etiam eget sapien nibh.

Nulla mi mi, Fig. 1 venenatis sed ipsum varius, volutpat euismod diam. Proin rutrum vel massa non gravida. Quisque tempor sem et dignissim rutrum. Lorem ipsum dolor sit amet, consectetur adipiscing elit. Morbi at justo vitae nulla elementum commodo eu id massa. In vitae diam ac augue semper tincidunt eu ut eros. Fusce fringilla erat porttitor lectus cursus, S1 Video vel sagittis arcu lobortis. Aliquam in enim semper, aliquam massa id, cursus neque. Praesent faucibus semper libero.

Figure 1. Figure Title first bold sentence Nulla mi mi, venenatis sed ipsum varius, volutpat euismod diam. Figure Caption Proin rutrum vel massa non gravida. Quisque tempor sem et dignissim rutrum. A: Lorem ipsum dolor sit amet. B: Consectetur adipiscing elit.

1. react
2. diffuse free particles
3. increment time by dt and go to 1

Results

Nulla mi mi, venenatis sed ipsum varius, Table 5 volutpat euismod diam. Proin rutrum vel massa non gravida. Quisque tempor sem et dignissim rutrum. Lorem ipsum dolor sit amet, consectetur adipiscing elit. Morbi at justo vitae nulla elementum commodo eu id massa. In vitae diam ac augue semper tincidunt eu ut eros. Fusce fringilla erat porttitor lectus cursus, vel sagittis arcu lobortis. Aliquam in enim semper, aliquam massa id, cursus neque. Praesent faucibus semper libero.

Table 5. Table caption Nulla mi mi, venenatis sed ipsum varius, volutpat euismod diam.

Heading1				**Heading2**			
*cell*1*row*1	cell2 row 1	cell3 row 1	cell4 row 1	cell5 row 1	cell6 row 1	cell7 row 1	cell8 row 1
*cell*1*row*2	cell2 row 2	cell3 row 2	cell4 row 2	cell5 row 2	cell6 row 2	cell7 row 2	cell8 row 2
*cell*1*row*3	cell2 row 3	cell3 row 3	cell4 row 3	cell5 row 3	cell6 row 3	cell7 row 3	cell8 row 3

Table notes Phasellus venenatis, tortor nec vestibulum mattis, massa tortor interdum felis, nec pellentesque metus tortor nec nisl. Ut ornare mauris tellus, vel dapibus arcu suscipit sed.

LOREM and IPSUM Nunc blandit a tortor.

Maecenas convallis mauris sit amet sem ultrices gravida. Etiam eget sapien nibh. Sed ac ipsum eget enim egestas ullamcorper nec euismod ligula. Curabitur fringilla pulvinar lectus consectetur pellentesque. Quisque augue sem, tincidunt sit amet feugiat eget, ullamcorper sed velit. Sed non aliquet felis. Lorem ipsum dolor sit amet, consectetur adipiscing elit. Mauris commodo justo ac dui pretium imperdiet. Sed suscipit iaculis mi at feugiat.

Sed ac quam id nisi malesuada congue.

Nulla mi mi, venenatis sed ipsum varius, volutpat euismod diam. Proin rutrum vel massa non gravida. Quisque tempor sem et dignissim rutrum. Lorem ipsum dolor sit amet, consectetur adipiscing elit. Morbi at justo vitae nulla elementum commodo eu id massa. In vitae diam ac augue semper tincidunt eu ut eros. Fusce fringilla erat porttitor lectus cursus, vel sagittis arcu lobortis. Aliquam in enim semper, aliquam massa id, cursus neque. Praesent faucibus semper libero.

Subsection 1

Nulla mi mi, venenatis sed ipsum varius, volutpat euismod diam. Proin rutrum vel massa non gravida. Quisque tempor sem et dignissim rutrum. Lorem ipsum dolor sit amet, consectetur adipiscing elit. Morbi at justo vitae nulla elementum commodo eu id massa. In vitae diam ac augue semper tincidunt eu ut eros. Fusce fringilla erat porttitor lectus cursus, vel sagittis arcu lobortis. Aliquam in enim semper, aliquam massa id, cursus neque. Praesent faucibus semper libero.

Subsection 2

3rd Level Heading. Nulla mi mi, venenatis sed ipsum varius, volutpat euismod diam. Proin rutrum vel massa non gravida. Quisque tempor sem et dignissim rutrum. Lorem ipsum dolor sit amet, consectetur adipiscing elit. Morbi at justo vitae nulla elementum commodo eu id massa. In vitae diam ac augue semper tincidunt eu ut eros.

Fusce fringilla erat porttitor lectus cursus, vel sagittis arcu lobortis. Aliquam in enim semper, aliquam massa id, cursus neque. Praesent faucibus semper libero.

Discussion

Nulla mi mi, venenatis sed ipsum varius, Table 5 volutpat euismod diam. Proin rutrum vel massa non gravida. Quisque tempor sem et dignissim rutrum. Lorem ipsum dolor sit amet, consectetur adipiscing elit. Morbi at justo vitae nulla elementum commodo eu id massa. In vitae diam ac augue semper tincidunt eu ut eros. Fusce fringilla erat porttitor lectus cursus, vel sagittis arcu lobortis. Aliquam in enim semper, aliquam massa id, cursus neque. Praesent faucibus semper libero.

LOREM and IPSUM Nunc blandit a tortor.

CO_2 Maecenas convallis mauris sit amet sem ultrices gravida. Etiam eget sapien nibh. Sed ac ipsum eget enim egestas ullamcorper nec euismod ligula. Curabitur fringilla pulvinar lectus consectetur pellentesque. Quisque augue sem, tincidunt sit amet feugiat eget, ullamcorper sed velit.

Sed non aliquet felis. Lorem ipsum dolor sit amet, consectetur adipiscing elit. Mauris commodo justo ac dui pretium imperdiet. Sed suscipit iaculis mi at feugiat. Ut neque ipsum, luctus id lacus ut, laoreet scelerisque urna. Phasellus venenatis, tortor nec vestibulum mattis, massa tortor interdum felis, nec pellentesque metus tortor nec nisl. Ut ornare mauris tellus, vel dapibus arcu suscipit sed. Nam condimentum sem eget mollis euismod. Nullam dui urna, gravida venenatis dui et, tincidunt sodales ex. Nunc est dui, sodales sed mauris nec, auctor sagittis leo. Aliquam tincidunt, ex in facilisis elementum, libero lectus luctus est, non vulputate nisl augue at dolor. For more information, see S1 Text.

Supporting Information

S1 Video

Bold the first sentence. Maecenas convallis mauris sit amet sem ultrices gravida. Etiam eget sapien nibh. Sed ac ipsum eget enim egestas ullamcorper nec euismod ligula. Curabitur fringilla pulvinar lectus consectetur pellentesque.

S1 Text

Lorem Ipsum. Maecenas convallis mauris sit amet sem ultrices gravida. Etiam eget sapien nibh. Sed ac ipsum eget enim egestas ullamcorper nec euismod ligula. Curabitur fringilla pulvinar lectus consectetur pellentesque.

S1 Fig

Lorem Ipsum. Maecenas convallis mauris sit amet sem ultrices gravida. Etiam eget sapien nibh. Sed ac ipsum eget enim egestas ullamcorper nec euismod ligula. Curabitur fringilla pulvinar lectus consectetur pellentesque.

S2 Fig

Lorem Ipsum. Maecenas convallis mauris sit amet sem ultrices gravida. Etiam eget sapien nibh. Sed ac ipsum eget enim egestas ullamcorper nec euismod ligula. Curabitur

fringilla pulvinar lectus consectetur pellentesque.

S1 Table

Lorem Ipsum. Maecenas convallis mauris sit amet sem ultrices gravida. Etiam eget sapien nibh. Sed ac ipsum eget enim egestas ullamcorper nec euismod ligula. Curabitur fringilla pulvinar lectus consectetur pellentesque.

Acknowledgments

Cras egestas velit mauris, eu mollis turpis pellentesque sit amet. Interdum et malesuada fames ac ante ipsum primis in faucibus. Nam id pretium nisi. Sed ac quam id nisi malesuada congue. Sed interdum aliquet augue, at pellentesque quam rhoncus vitae.

References

1. Sebastian Raschka. Python Machine Learning Essentials. Packt Publishing; 2015.
2. Cam Davidson-Pilon. Quickstart – lifelines 0.8.0.1 documentation; 2016 (accessed 14 Jan 2016). http://lifelines.readthedocs.org/en/latest/Quickstart.html.
3. Van Poucke S, Zhang Z, Schmitz M, Vukicevic M, Laenen MV, Celi LA, et al. Scalable predictive analysis in critically ill patients using a visual open data analysis platform. PLoS ONE. 2016;11(1). Cited By 0. Available from: http://www.scopus.com/inward/record.url?eid=2-s2.0-84953931466&partnerID=40&md5=7a0cad7137c03146e4b75f3295f84cc6.
4. National Cancer Institute, the Surveillance, Epidemiology, and End Results Program. About the SEER Program - SEER; 2016 (accessed 14 Jan 2016). http://seer.cancer.gov/about.
5. Shin, Hyunjung and Nam, Yonghyun; ISCB Asia. A coupling approach of a predictor and a descriptor for breast cancer prognosis [Article; Proceedings Paper]. BMC MEDICAL GENOMICS. 2014 MAY 8;7(1). 3rd Annual Translational Bioinformatics Conference (TBC) / ISCB-Asia, Seoul, SOUTH KOREA, OCT 02-04, 2013.
6. Zolbanin, Hamed Majidi and Delen, Dursun and Zadeh, Amir Hassan. Predicting overall survivability in comorbidity of cancers: A data mining approach [Article]. DECISION SUPPORT SYSTEMS. 2015 JUN;74:150–161.
7. Gordon L, Olshen RA. Tree-structured survival analysis. Cancer Treatment Reports. 1985;69(10):1065–1068. Cited By 97. Available from: http://www.scopus.com/inward/record.url?eid=2-s2.0-0021875130&partnerID=40&md5=9e112ed840960f801b6260b23bf6811d.
8. Bou-Hamad I, Larocque D, Ben-Ameur H. A review of survival trees. Statistics Surveys. 2011;5:44–71. Cited By 15. Available from: http://www.scopus.com/inward/record.url?eid=2-s2.0-84857308440&partnerID=40&md5=f8af82017ade68e335fd258c6857bf49.

9. Ishwaran H, Kogalur UB. Consistency of random survival forests. Statistics and Probability Letters. 2010;80(13-14):1056–1064. Cited By 26. Available from: http://www.scopus.com/inward/record.url?eid=2-s2.0-77953020220&partnerID=40&md5=1e4478c51150f0159fdc6c1cb631968b.

10. National Cancer Institute, the Surveillance, Epidemiology, and End Results Program. Documentation for ASCII Text Data Files - SEER Datasets; 2016 (accessed 15 Jan 2016). http://seer.cancer.gov/data/documentation.html.

11. Michael Bowles. Machine Learning in Python: Essential Techniques for Predictive Analysis. Wiley; 2015.

12. Allen Downey. Think Stats. O'Reilly Media; 2014.

13. United States Census Bureau. 2010 FIPS Code Files for Counties - Geography - U.S. Census Bureau; 2016 (accessed 18 Jan 2016). https://www.census.gov/geo/reference/codes/cou.html.

14. Google Developers. The Google Maps Geocoding API — Google Maps Geocoding API — Google Developers; 2016 (accessed 18 Jan 2016). https://developers.google.com/maps/documentation/geocoding/intro.

15. Google Developers. The Google Maps Elevation API — Google Maps Elevation API — Google Developers; 2016 (accessed 18 Jan 2016). https://developers.google.com/maps/documentation/elevation/intro?hl=en.

16. IOBS. Supplemental Material — PAPERDATA; 2016 (accessed 18 Jan 2016. https://github.com/doolingdavid/PAPERDATA.git.