

**Design and Implementation of Analytics System**

***Predicting Breast Cancer Survival Period and Recommending Treatments***

***Project 02***

*Rishu Gandhi*

*Fady Smouni*

**School of Graduate Professional Studies**

MPS Data Analytics (DAAN)

DAAN 888 – Design and Implementation of Analytics System

Fall, 2021

Supervisor: Robin Qiu, Ph.D.

# Document Control

## Work carried out by:

|  |  |  |
| --- | --- | --- |
| **Members** | **Email Address** | **Task description** |
| Rishu Gandhi | Rmg5430@psu.edu | Data Collection & Cleaning, Model building, Feature Engineering, Bayesian Network exploration |
| Fady Smouni | Ffs5130@psu.edu | Data exploration, Scientific survey, RNN Model building, Feature Engineering, API building, Bayesian Network exploration |

## Revision Sheet

|  |  |  |
| --- | --- | --- |
| **Release No.** | **Date** | **Revision Description** |
| 1 | 8/29/2021 | Worked on finding datasets for Breast Cancer |
| 2 | 9/5/2021 | Found research articles and explored the dataset |
| 3 | 9/12/2021 | Worked on data collection & exploration, created dictionary table |
| 4 | 9/19/2021 | Worked on data cleaning, removing duplicates & missing values, outliers. Etc. |
| 5 | 9/24/2021 | Worked on defining which model to be used |
| 6 | 10/3/2021 | Worked on RNN & Regression Analysis |
| 7 | 10/10/2021 | Created models for RNN & Regression |
| 8 | 10/17/2021 | Optimized models for better accuracy |
| 9 | 11/14/2021 | Modified used dataset, updated RNN, started work on Bayesian Network |
| 10 | 11/21/2021 | Updated Visualization for RNN and updated Bayesian Network progress, added API description for model deployment, briefly discussed results, limitations, and future work. |
| 11 | 12/6/2021 | Final Revision |

**TABLE OF CONTENTS**

Contents

[Document Control 1](#_Toc89685428)

[1. Work carried out by: 1](#_Toc89685429)

[2. Revision Sheet 1](#_Toc89685430)

[Project Proposal 1](#_Toc89685431)

[1. Project title 1](#_Toc89685432)

[2. Background and Challenges 1](#_Toc89685433)

[3. Project Objectives 1](#_Toc89685434)

[4. Feasibility Studies 2](#_Toc89685435)

[5. Methodology (Data Analytics System) 3](#_Toc89685436)

[6. Deliverables 3](#_Toc89685437)

[7. Planned Steps: 3](#_Toc89685438)

[8. Importance and Impacts 3](#_Toc89685439)

[9. Required Technologies 3](#_Toc89685440)

[10. References 4](#_Toc89685441)

[Data Collection, Data Cleaning 5](#_Toc89685442)

[1. Data Source 5](#_Toc89685443)

[2. Data Collection 5](#_Toc89685444)

[3. Data Storage 5](#_Toc89685445)

[4. Data Pertinence 5](#_Toc89685446)

[5. Data Dictionary 6](#_Toc89685447)

[6. Data Cleaning 9](#_Toc89685448)

[Feature Engineering and Methodology 12](#_Toc89685449)

[1. Feature Engineering 12](#_Toc89685450)

[2. Deep Neural Networks 15](#_Toc89685451)

[3. Regression Models 19](#_Toc89685452)

[4. Bayesian Network Exploration 23](#_Toc89685453)

[Model Deployment 26](#_Toc89685454)

[1. Overview 26](#_Toc89685455)

[2. Model Saving 26](#_Toc89685456)

[3. API Guide 26](#_Toc89685457)

[4. API Example 27](#_Toc89685458)

[Results, Limitation, Future Work 28](#_Toc89685459)

[1. Discussion of Results 28](#_Toc89685460)

[2. Limitations 28](#_Toc89685461)

[3. Future Work 28](#_Toc89685462)

# Project Proposal

## Project title

1. Predicting Breast Cancer Survival Period and Recommending Treatments

## Background and Challenges

1. Breast cancer is a disease in which breast cells grow exponentially and some of them can be deadly. According to the American cancer Society, 13% of women in the U.S. are at risk of developing breast cancer during their lifetime. Bringing breast cancer to be the second leading cause of death related to cancer in women, topped only by lung cancer [10].
2. Using machine learning models can help identify survival chances of a patient. There have been many studies performed on tumor related features like tumor size and the age it was diagnosed to predict survival chances. We wish to implement machine learning techniques to predict the chance of patients. Based on past cases, we can classify various groups of patients according to their various demographic variables, and medical data related to their cancer diagnosis. The challenge lays in understanding the various features and parameters that are encompassed within this medical data. As we are not medical professionals and do not have access to a medical professional that could assist us with getting a better understanding of the variables and maybe understanding which ones are irrelevant or more important from a professional in the subject matter

## Project Objectives

1. We aim to perform classification based on various parameters to determine a survival range for the patient.
2. We would also like to try and perform a regression analysis to see If it is possible to predict how long a patient will survive in months.
3. Try to explore the use of Bayesian Networks in the understanding which treatment(s) could most likely prolong breast cancer patients’ survival time at their different survival periods based on a survival causal analysis.

## Feasibility Studies

* Keywords: Breast Cancer, SEER, NIH, RNN, Survival, Deep Learning
* The main piece of work that is highly similar to our project is an unpublished paper “Deep Learning and Prediction of Survival Period for Breast Cancer Patients” by Shreyesh Doppalapudi, Hui Yang, Jerome Jourquin, and Robin G. Qiu. This paper lays the ground for a proof of concept that shows there is a basis for developing models that can assist medical professionals in planning the patient’s journey. This paper shows that deep neural networks perform much better than traditional machine learning models (random forest, linear regression…).
* They are various papers and project that have aimed to bridge the gap between the medical field and the machine learning community. One paper that tried to do that by Li et al [1]. Have shown that when predicting breast cancer 5-year survival using traditional machine learning models we do not get significantly better results than traditional statistical methods. This paper also brings up issues related to the lack of available data. The SEER program has been a great answer to that issue and having access to that data will allow us to circumvent those limitations.
* Montazeri et al [2]. shows that using Trees Random Forest model represents the best route when using traditional machine learning models with a 96% accuracy. However, these results are not in line with the majority of other projects and research papers, which drives us to question the methods and dataset used to achieve such a high accuracy with a traditional machine learning method.
* Ganggayah et al [3]. Looks at determining the important factors of for the survival of breast cancer patients using machine learning techniques. They used various algorithms in their research which all provided relatively close results with decision tree at the low end (79.8%) and random forest at the high end (82.7%). They were able to assess that cancer stage classification, tumor size, number of total axillary lymph nodes removed, number of positive lymph nodes, types of primary treatment, and methods of diagnosis were all important factors in the prediction the survival of breast cancer patients.
* Hareendran et al [5]. Show that a combination of naïve bayes and deep neural networks are much better predictors than classical methods when looking at predicting survival rate for patients.
* Initially, our project will be highly similar to the “Deep Learning and Prediction of Survival Period for Breast Cancer Patients” in order to confirm the results from that research since it uses the same dataset. However, we will take our research further by trying to analyze causal inference based on clusters, survival causal analysis will be applied to understanding which treatment(s) could most likely prolong breast cancer patients’ survival time at their different survival periods.

## Methodology (Data Analytics System)

* As stated above, the dataset is used in the unpublished paper “Deep Learning and Prediction of Survival Period for Breast Cancer Patients” by Shreyesh Doppalapudi, Hui Yang, Jerome Jourquin, and Robin G. Qiu.

1. Explain briefly the data analytics workflow and pipeline to solve your problem
2. Clearly list the task that will be completed in the following weeks.
3. How your methodology proves/disapproves your project objectives.

## Deliverables

* Commented Final Code
* Used datasets
* Saved model
* API for user to use saved model for prediction
* Slides for presentation
* This complete report

## Planned Steps:

* **Stage I:** Understand the data

1. Work on understanding all the different attributes related to the data
2. Clean the data (i.e., check for missing values, duplicates, preprocess features, correlation…)

* Stage II: Try to improve on the results provided by the paper that we are building our research on of “Deep Learning and Prediction of Survival Period for Breast Cancer Patients” by Shreyesh Doppalapudi, Hui Yang, Jerome Jourquin, and Robin G. Qiu.

1. Binary (e.g., x1<=6, x2>6) classification to predict clusters
2. The exact survival time (in months) regression models based on the result in Step1

* **Stage III:** Exploring Bayesian network approach to analyze causal inference based on clusters, survival causal analysis will be applied to understanding which treatment(s) could most likely prolong breast cancer patients’ survival time at their different survival periods

## Importance and Impacts

* Research shows that breast cancer can affect 1 in 8 U.S. women. That is why it is important to do everything we can to add value to the immense research that is done to fight this disease. Significant progress has been achieved over the past decades, bringing survival rates higher and allowing patients to live longer. Through our research, we aim to help drive this progress further by providing a clearer picture of patient survival and potentially start a conversation about life extension in relation to the various availale treatments.

## Required Technologies

* Languages: Python
* Software: Jupyter\_Notebook, Anaconda
* Libraries: pandas, NumPy, matplotlib, TensorFlow, sklearn, statsmodels, seaborn, causalnex

## References

1. Li, J. L., Zhou, Z. Z., Dong, J. D., Fu, Y. F., Li, Y. L., Luan, Z. L., & Peng, X. P. (2021, April 16). Predicting breast cancer 5-year survival using machine learning: A systematic review. PubMed Central (PMC). <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8051758/>​
2. Montazeri, M. M., Montazeri, M. M., Montazeri, M. M., & Beigzadeh, A. B. (2016). Machine learning models in breast cancer survival prediction. PubMed. <https://pubmed.ncbi.nlm.nih.gov/26409558/>​
3. Ganggayah, M. D., Taib, N. D., Har, Y. C., Lio, P., & Dhillon, S. K. (2019, March 22). Predicting factors for survival of breast cancer patients using machine learning techniques. BMC Medical Informatics and Decision Making. <https://bmcmedinformdecismak.biomedcentral.com/articles/10.1186/s12911-019-0801-4>​
4. Mihaylov, I., Nisheva, M., & Vassilev, D. (2019). Application of Machine Learning Models for Survival Prognosis in Breast Cancer Studies †. MDPI. <https://www.mdpi.com/2078-2489/10/3/93>​
5. Hareendran, A. S., Chandra, V., Prasad, S. R., & Dhanya, S. (2020, July 14). Deep Learning Strategies for Survival Prediction in Prophylactic Resec. SpringerLink. <https://link.springer.com/chapter/10.1007/978-3-030-53956-6_53>​
6. Kourou, K. K., Exarchos, T. P., Exarchos, K. P., Karamouzis, M. V., & Fotiadis, D. I. (2015, January 1). Machine learning applications in cancer prognosis and prediction. ScienceDirect. <https://www.sciencedirect.com/science/article/pii/S2001037014000464>
7. “causalnex”. Accessed November 25, 2021. <https://causalnex.readthedocs.io/en/latest/04_user_guide/04_user_guide.html>
8. Rosebrock, A. (2021, April 17). Keras Tutorial: How to get started with Keras, Deep Learning, and Python. PyImageSearch. <https://www.pyimagesearch.com/2018/09/10/keras-tutorial-how-to-get-started-with-keras-deep-learning-and-python/>
9. SEER Program, National Cancer Institute (NCI), 'SEER Incidence Data, 1975-2016', Available: <https://seer.cancer.gov/data/>
10. American Cancer Society (2021, May 7), ‘How Common Is Breast Cancer?’. <https://www.cancer.org/cancer/breast-cancer/about/how-common-is-breast-cancer.html>

# Data Collection, Data Cleaning

## Data Source

The Breast Cancer data was collected from SEER program registry [9]. The Surveillance, Epidemiology, and End Results (SEER) program provides various cancer statistics for research purposes. Their data is used by various researchers and scientists to research the cancer causes and predict treatments. The SEER program collects cancer related data from various locations and sources throughout the Unites States. This geographic statistic helps to learn about cancer situations of the Unites States’s overall population.

## Data Collection

The dataset was collected through a request form on the SEER website. This allowed us to receive the data as a text file. Please note that this task was performed by our project supervisor Robin Qiu, Ph.D.

## Data Storage

Our data is stored in a csv file after being extracted from the text file provided through the SEER program.

## Data Pertinence

This dataset is very relevant to our problem statements; it includes statistics on tumor size, demographic information, treatments acquired, and diagnostic information. All these features help to identify the important features more susceptive to cancer survivability.

## Data Dictionary

|  |  |  |  |
| --- | --- | --- | --- |
| 1. **Attributes** | **Definition** | **Data Type** | **Data Values** |
| MAR\_STAT | Marital Status | Nominal | Code: 1 - 9 |
| SEX | ER Status Record Breast Cancer (1900+) | Binary | Code: 1 - 2 |
| SEQ\_NUM | SEQUENCE NUMBER--CENTRAL | Nominal | Code: 00 - 99 |
| PRIMSITE | Primary Site | Nominal | - |
| LATERAL | Laterality | Nominal | Code: 0 - 9 |
| GRADE | Grade | Nominal | Code: 1 - 9 |
| SURGPRIF | X Summ—Surg Prim Site | Nominal | Code: 00 - 99 |
| SURGSITF | RX Summ—Surg Oth Reg/Dis | Nominal | Code: 0 - 9 |
| NO\_SURG | REASON FOR NO SURGERY | Nominal | Code: 0 - 9 |
| AGE\_1REC | AGE RECODE <1 YEAR OLDS | Nominal | Code: 00 - 99 |
| BEHTREND | BEHAVIOR RECODE FOR ANALYSIS | Nominal | Code: 0 - 6 |
| RAC\_RECA | Race recode (White, Black, Other) | Nominal | Code: 1 - 9 |
| CODPUB | Cause of Death to SEER site recode | Nominal | - |
| CODPUBKM | COD TO SITE REC KM | Nominal | - |
| STAT\_REC | VITAL STATUS RECODE | Binary | Code: 0 - 1 |
| ERSTATUS | ER Status Record Breast Cancer (1900+) | Nominal | Code: 1 - 9 |
| PRSTATUS | PR Status Recode Breast Cancer (1900+) | Nominal | Code: 1 - 9 |
| SRV\_TIME\_MON | Survival Months | Discrete |  |
| INSREC\_PUB | INSURANCE RECODE (2007+) | Nominal | Code: 1 - 5 |
| ADJTM\_6VALUE | Breast Adjusted AJCC 6th T (1988-2015) | Nominal | - |
| ADJNM\_6VALUE | Breast Adjusted AJCC 6th N (1988-2015) | Nominal | - |
| ADJM\_6VALUE | Breast Adjusted AJCC 6th M (1988-2015) | Nominal | - |
| ADJAJCCSTG | BREAST ADJUSTED AJCC 6TH STAGE (1988-2015) | Nominal | - |
| her2 | Derived HER2 Recode (2010+) | Nominal | Code: 1 - 9 |
| brst\_sub | Breast Subtype (2010+) | Nominal | Code: 1 - 9 |
| MALIGCOUNT | Total Number of In Situ/malignant Tumors for Patient | Discrete | - |
| BENBORDCOUNT | Total Number of Benign/Borderline Tumors for Patient | Discrete | - |
| RADIATION | Type of Radiation | Nominal | Code: 0 - 8 |
| RAD\_SURG\_SEQ | Sequence of Surgery & Radiation Therapy | Nominal | Code: 0 - 9 |
| CHEMO | Chemotherapy | Binary | Code: 0 - 1 |

## Data Cleaning

* + **Missing Values**

Our initial dataset contains 840,666 rows and 30 columns. We first start by checking for missing values. We have decided to not do any data imputing for missing values. That is because we are dealing with medical data that differs from one patient to the other and has an important impact on how these patients’ health evolves.

**Table

Description automatically generated**

As we can see above, there are a number of missing values within the dataset. Since we decided to not impute any values in order to protect the integrity of the medical data, we decided to drop any rows with missing values at it will impact how our model performs.

* + **Duplicates**

In terms of duplicates, our dataset did not have any duplicates. So no duplicate elimination was necessary.

Graphical user interface, text, application

Description automatically generated

* + **Skewness and Type conversions**

There are many attributes with nominal categories. These data attributes are kept with original data type for measuring multicollinearity purpose. They will be changed later during model exploration.

* + **Correlation Check**

We then look into correlation of the data. We can clearly see below that there is significant multicollinearity for certain attributes. However, because we are dealing with medical data, and we believe having these attributes bring important information to the tale that is needed by our model to make better predictions.

Graphical user interface

Description automatically generated with medium confidence

* + **Histograms to Check for Skewness**

We can clearly see in the image below trough the histogram representations that we have a certain amount of skewness in many of the attributes that we are using within our dataset. However, this is to be expected, as we are keeping our dataset close to its original state to maintain the medical importance of many of its features.

A picture containing window, crossword puzzle

Description automatically generated

# Feature Engineering and Methodology

In this project, we are going to perform classification using deep neural network. Our goal is to be able to classify patient into survivability groups based on their survival time. We will then try and predict the exact survival of the patient in terms of years using regression models. Finally, we are going to explore the potential applications of Bayesian networks in helping us understand the impact of each treatment type and path in the survivability of the patient.

## Feature Engineering

In order to extract the most performance of our model, we had to do some data preprocessing, also known as feature engineering.

We start by dropping all male participants. This is due to the fact that males represent a very small portion of the entire population for this dataset and therefore will negatively affect our model. By getting rid of the male records, we are also able to drop the Sex category as it is no longer needed since we only have females left.

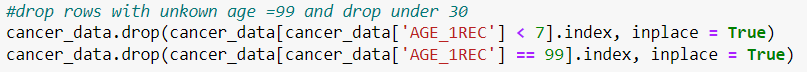
Text

Description automatically generated

We then decided to change our no\_surg column to a Boolean representation by grouping all no surgeries together regardless of the reason they were not performed. This allows us to give our model less granularity to shift trough.



We also drop patients whose age is unknow or below 30 years old. This is due to the fact that patients below 30 years old represent a very small portion of the dataset and create unnecessary complexity for our model.

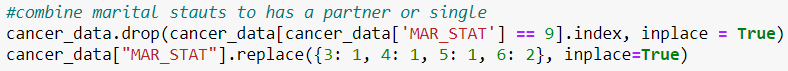


We also combine the race groups of other and unknow into one category. As it is unnecessary to have two other groups and an unknow.

Text

Description automatically generated with medium confidence

We also decide to simplify the marital status by combining all the categories into two. Has a partner or single.



The dataset from SEER has patients that died of various reasons. We decide to only keep patients that died of breast cancer or are still alive. After filter all those other patients, we also drop both columns that provide that information as we do not want to tell the model that a patient is dead or alive since we cannot provide that information when using it in the real world.

A picture containing icon

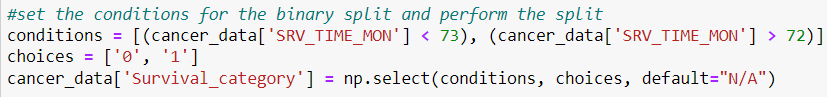
Description automatically generated

We combine insured and insured/no specifics as they both are just insured.



We also drop various unknow categories from other columns.

Finally, we are able to add a new column to represent our categorical split for our classification. We slit patients based on survival time. Patients that survived over 6 years and patients that survive for 6 years or less.



Then we can drop the survival time in months column as it is no longer pertinent for us.

Text

Description automatically generated with medium confidence

Then we one hot encodes the variables within the dataset as they all categorical in nature. Except for the Survival category as it is what we want to predict. By one hot encoding the variables, we get rid of any numerical order the model might give to different categories and therefore giving them more importance.

Chart, scatter chart

Description automatically generated

We then split our data into labels and data.

Company name

Description automatically generated with medium confidence

Then to be able to fit the data into the RNN model, we need to turn it into an array. But we do not need to do any scaling as all our variables are binary in nature after the one hot encoding.

A picture containing text

Description automatically generated

We then reshape the data array as it needs to be 3 dimensional for the LSTM RNN model.



Now to prepare our training and testing splits, we use an 80/20 split for training and testing and a 90/10 split for training and validation.

We also need to vectorize our label training, test, and validation sets so that we can use them for prediction.

Text

Description automatically generated

## Deep Neural Networks

For our classification implementation, we base our process on the results provided by Doppalapudi et al., on their unpublished paper “Deep Learning and Prediction of Survival Period for Breast Cancer Patients”. It is clear from their research that RNN is the best structure to use for our classification model. In our initial approach, we decided to use a similar classification to what was done by Doppalapudi et al., however, there was clearly a limit to the results that can be attained with that less than 5 years and more than 5 years classification. We ran multiple tests on this type of classification. However, we were not able to improve upon the 78% accuracy that was achieved on the original paper.

Struggling to make any progress using various models and hyper parameter tunning, we decided to have a deeper look into the predictions of the model. We ran our model on the entire dataset and found out that when comparing predicted with actual, the model had a hard time distinguishing between patients that survive 5 years and patients that survive 6 years. This represented a major turning point for us. We decided to change our classification to less than 6 years and more than 6 years. This represented a major break in our efforts to improve on the accuracy of our models. Using this classification structure, we were able to achieve a 90% accuracy.

Our RNN model was structured as 2 hidden layers of size 64 and 32 respectively. A relu activation function after each hidden layer and a 0.1 dropout. We had a sigmoid activation layer as this was a binary problem. We used the Adam optimizer with a learning rate of 0.0001 and the binary\_crossentropy loss function.

Text

Description automatically generated

We also ran a simple neural network as a benchmark to have an additional comparison point for our results. This model had 5 hidden layers of size 600, 300, 100, 50, 20 respectively. With a relu activation function and a dropout of 0.1 after the first two hidden layers. The model achieved 87% accuracy. Which represents a baseline that shows us that the RNN model performs better.

Text

Description automatically generated

As we saw above, these original models were trained using a train test split of 80% for training. In order to make sure that randomness is not a factor in the performance of our model, we decide to run a 5-fold cross validation of our model. This cross validation allows us train and test our model using all of our dataset and to make sure that our initial model accuracy was not just the result of the test dataset having easy to predict values.

The results of the cross validation confirm our model performance. We have an average accuracy of 90.29% (+/- 0.14%). We show in the table below the performance of each fold.

|  |  |
| --- | --- |
| Fold | Accuracy (%) |
| 1 | 90.26% |
| 2 | 90.33% |
| 3 | 90.50% |
| 4 | 90.30% |
| 5 | 90.07% |

Text

Description automatically generated

Seeing that our model performs at the same relative accuracy across all the folds, we are able to train our model on the entire dataset without the need for a testing set. This allows us to use 20% more data to train our model. That is the model that we will be saving and using for prediction within our API.

Graphical user interface, text

Description automatically generated

Finally, we can evaluate our RNN model. We will use the model that was training on the 80/20 split for graphical representation.

Text

Description automatically generated

Graphical user interface

Description automatically generated

Table

Description automatically generated with medium confidence

As we can see above, our model performs very well. We have an accuracy of 90%. Our precision, recall, and f1-score are all within relatively good margins.

## Regression Models

The regression model uses the same dataset as our RNN model except for the fact that it has a survival time in years column that we predict on.

Text

Description automatically generated

Model 1- Patients with more than 6 years of survival

The PCA analysis were performed for both models to reduce dimensionality in the attributes

Graphical user interface, text, application, email

Description automatically generated

After preparing the PCA Analysis, Logistic regression was performed as below.

Table

Description automatically generated

The accuracy was found to be 24% with the newly modified data, which is very poor.

Model 2- Patients with more than 6 years of survival

Graphical user interface, text, application, email

Description automatically generated

Table

Description automatically generated

The accuracy is found to be 53%, which is slightly better than the model for less than 6 years of survival.

Overall logistic regression did not perform well with 24% and 53% accuracy for Model 1 and Model 2 respectively. This shows that there may not a linearity between the attributes in predicting the survival years. Its is more complicated than that. It also is underperformed due to multiple non-linear boundaries within attributes.

## Bayesian Network Exploration

Finally, we look into exploring Bayesian Networks to interpret the impact that certain treatment courses have on patients. Our goal here is not to build a perfect model. We are only exploring the possibilities and starting a conversation regarding this.

For this, we restrict our dataset to only these variables: "SRV\_TIME\_YR", "NO\_SURG", "RADIATION", "RAD\_SURG\_SEQ", "CHEMO"

When using Bayesian network, we can try and get conditional probabilistic distributions. This allows us to set certain condition that will provide a certain outcome and see the probability of various categories of that outcome happening based on a set of events.

In order to use Bayesian networks, we need to set our models edges and structure. For our case, we do not have an expert in breast cancer. So, we assume the connection to the best of our knowledge.

Our structure is as follows:

Text, letter

Description automatically generated

NO\_SURG

SRV\_TIM\_TR

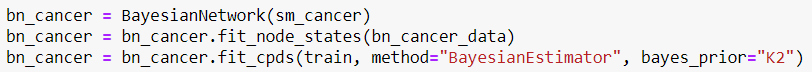
CHEMO

RADIATION\_SURG\_SEQ

RADIATION

After we set up our structure, we can train our model using an 80/20 split.





We perform an area under the curve test. The results are not perfect, but they are acceptable for our purpose which is the exploration of this method.

Text

Description automatically generated with medium confidence

Now we can look at the probabilistic distribution of our model. We can basically choose certain combinations of event to see the survival probability by year of our patients.

For example, we can look at the survival probabilistic distributions of a patient that has had chemotherapy, has had a surgery, had Beam radiation, and had radiation before surgery.

Graphical user interface, text, application, email

Description automatically generated

As we can see above, patients have a higher probability of surviving past the first two years with this combination of medical treatments.

We can have another look at a different treatment combination, had chemotherapy, did not have a surgery, had Beam radiation.

Graphical user interface, text

Description automatically generated

As we can see here, the fact that the patient did not have a surgery significantly impacts their probabilistic distribution in terms of survivability in years. The patient has less chances to survive past year 3 compared to when they had the surgery.

We can clearly see that Bayesian networks open the opportunity for treatment sequencing that would allow us to better understand which type of treatments and treatment combinations best work for extending the survivability of patients.

We can extend our model by adding other variables such as the persons age, the number of tumors they have.

# Model Deployment

## Overview

For our model deployment, we will use a simple API python script that will allow the user to input a csv file that contains all the necessary information needed about the patient as described below in the API guide section. This API will allow the user to also load our model to use. We allow the user to choose which model they want to load, in case we come up with new models in the future. After the user uploads the necessary csv file and model file, we will predict using our model and output the survival category that the patient falls within.

Our API python script is inspired and based on the example provided by Rosebrock in his article Keras Tutorial: How to get started with Keras, Deep Learning, and Python [8].

## Model Saving

We save our fully trained RNN\_model using the code below.



## API Guide

This API guide is also provided as a Read me file named “Installation.txt” within the file package for this project.

We provide the following API script "Python\_Predict" for users to test our RNN models on their own patient data.

We currently only support imputing one patient at a time.

The input needs to be a csv file that contains the following information structured over two rows, with the first row being the title of the rows

and second row containing the data.

The data that needs to be provided is MAR\_STAT, SEQ\_NUM, PRIMSITE, LATERAL, GRADE, SURGPRIF, SURGSITF, NO\_SURG, AGE\_1REC, BEHTREND, RAC\_RECA,

STAT\_REC, ERSTATUS, PRSTATUS, INSREC\_PUB, ADJTM\_6VALUE, ADJNM\_6VALUE, ADJM\_6VALUE, ADJAJCCSTG, her2, brst\_sub, MALIGCOUNT, BENBORDCOUNT,

RADIATION, RAD\_SURG\_SEQ, CHEMO

Please review the SEER\_TextData\_FileDescription file that is provided in the overall zipped folder. It will have a description of what each one of

these parameters represents and what kind of data needs to be imputed for each one.

Please note that for "PRIMSITE", do not input the first letter. Just the last 3 numbers.

The API takes two arguments, the data argument which is the path to the data file the user wants to use.

We also take the model argument which is the path to the model the user wants to use (For our case, that would be the RNN\_model\_Final)

We have added this functionality to allow the user the ability to use any future models we might publish.

Here is an example of how the user can call this API

(PyLab3) C:\Users\fadyn\Desktop\...\Final>python Python\_Predict.py --data Cancer\_data\_API\_Final.csv --model RNN\_model

You will find the models that can be used in the same zip file as this Read me file.

We also put 1 patient sample data to be used for prediction in this zip file under file name Cancer\_data\_API\_Final.

If you have any questions about this API, please contact Fady Smouni at ffs5130@psu.edu or Rishu Gandhi at rmg5430@psu.edu

## API Example





# Results, Limitation, Future Work

## Discussion of Results

For RNN, we believe that we have a very good model with an accuracy of 90%. This is a clear improvement on the previous models that were run on the original paper for this project. Our model is able to help classify patients into two categories. Patients are put into a 6 year or less survival period and a more than 6 years survival period.

As we saw, the Bayesian network is very promising as a model to see causal probabilistic distribution of treatments in relation to survival time in years.

## Limitations

The construction and implementation of BNN model with our dataset has been very challenging for us. We wish to spend more time exploring and optimizing the BNN model capabilities. One of the limitations is our understanding of how Bayesian networks work and their mathematical implications.

For regression, the limitation was the type of data attributes. We had to use one-hot coding technique to convert most of the attributes and use logistic regression instead, which did not perform well. We were not able to display correlations within attributes effectively using regression model. We applied PCA analysis to reduce dimensionality of the data, but it did not help unfortunately. The RMSE and MAE of linear regression showed that the prediction error was about 13 months in predicting survival months of a patient.

For RNN, one of the limitations we have currently resides with the fact that we are unable to go into more granularity and predict the survivability of patients on a monthly or yearly basis. This type of prediction seems to be too complex for the various models we tried and does not provide adequate accuracy.

## Future Work

Future work that can be undertaken for this project would work on getting new variables into the dataset to help increase the ability of the model to predict on a more granular level and maybe predict the survivability of individuals on a monthly or yearly aspect.

We would also want to take the BNN part further and build a more robust network that would allow us to see which kind of treatment works better for certain categories of individuals.