

APPLICATION OF MACHINE LEARNING IN  
PREDICTING OVARIAN CANCER  
SURVIVABILITY

By

VEDIKA HARISHKUMAR DENGADA

Bachelor of Science in Production Engineering

Fr. Conceicao Rodrigues College of Engineering

Mumbai, Maharashtra

2011

Submitted to the Faculty of the  
Graduate College of the  
Oklahoma State University  
in partial fulfillment of  
the requirements for  
the Degree of  
MASTER OF SCIENCE  
July, 2015

APPLICATION OF MACHINE LEARNING IN  
PREDICTING OVARIAN CANCER  
SURVIVABILITY

Thesis Approved:

Dr. Tieming Liu

---

Advisor

Dr. Dursun Delen

---

Co-Advisor

Dr. David Pratt

---

Committee Member

## ACKNOWLEDGEMENTS

Working on the thesis has been a unique experience. It was always a constant learning and improving process, which was made possible by support and expert guidance.

I take this opportunity to express my gratitude to Dr. Liu, without whom I would not have taken up this opportunity. Dr. Liu has truly made a difference in my life. His persistence, understanding, kindness and patience have been a major support for the completion of my research. Dr. Liu's keen interest at every stage of the research inspired me in completing the research within schedule.

I wish to express my sincere thanks to Dr. Delen for keeping me motivated with his encouragement. He provided me with direction, in solving technically complex problems and helped me face challenges with ease. I am extremely thankful and indebted to him for sharing his expertise and for the sincere and valuable guidance he extended to me. This thesis would not have been possible without his immense support.

Sincere thanks to Dr. Pratt for inspiring me to look into the details and improve. I would also like to extend my sincere thanks to Dr. William Paiva who gave me an opportunity to work on the Center for Health Systems Innovation (CHSI)'s Cerner Health Facts® data warehouse and use a portion of it for my research.

Finally, I would like to express my gratitude and sincere thanks towards my parents Mr. Harishkumar & Mrs. Nirmala Dengada. Without their trust, vision, love & support nothing would have been possible. Sincere thanks to my elder sister Mrs. Amrapali Dolare, my younger sister Ms. Krishna Dengada and my best friend and well-wisher Mr. Aditya Shringarpure.

Name: VEDIKA H DENGADA

Date of Degree: JULY, 2015

Title of Study: APPLICATION OF MACHINE LEARNING IN PREDICTING  
OVARIAN CANCER SURVIVABILITY

Major Field: INDUSTRIAL ENGINEERING AND MANAGEMENT

Abstract: This study applies machine learning techniques for predicting ovarian cancer survivability using Cerner Health facts data. Specifically, the study uses three popular data mining techniques: Neural network- MLP, Neural network-RBF and Decision trees. Using 10-fold cross validation is used in all the techniques to minimize the overfitting of models. Based on the descriptive statistics this study finds similar patterns that are observed in studies conducted on SEER cancer data. The aggregated results indicate that balanced technique using neural network multilayer perceptron in IBM SPSS modeler performed the best with a classification accuracy of 97.71% which is better than any other model compared in the study. The second best is using unbalanced data on neural network radial basis function with a classification accuracy of 96.18%. The neural network with radial basis function comes out as the worst with a classification accuracy of 67.80% even with a balanced dataset. This signifies given a set of parameters used in the study like: admission source, race of the patient, census division and so on the neural network using multilayer perceptron will predict the outcome of survival of the patient with 97.71% accuracy. In addition to the prediction model this study also found important factors in order to have a better insight into the relative contribution of the variables to predict survivability.

## TABLE OF CONTENTS

Chapter	Page
I. INTRODUCTION.....	1
1.1 About Ovarian Cancer .....	3
1.2 Signs and Symptoms.....	4
1.3 Type of Ovarian Cancer.....	5
1.4 Staging of Cancer.....	5
1.5 Causes and Risk Factors .....	6
1.6 Diagnosis.....	7
1.7 Problem Statement .....	9
1.8 Research Objectives.....	9
II. LITERATURE REVIEW .....	11
III. METHODOLOGY .....	16
3.1 Introduction.....	16
3.2 CRISP- DM Approach.....	18
3.2.1 Business Understanding.....	19
3.2.2 Data Understanding .....	20
3.2.3 Data Preparation.....	20
3.2.4 Modeling.....	23
3.2.5 Evaluation .....	23
3.2.6 Deployment .....	23
3.3 Descriptive Profiling.....	23
3.4 Predictive Modeling.....	24
3.4.1 Balancing Technique .....	24
3.4.2 Predictive Modeling Classification Methods.....	25
i. Decision Trees .....	26
ii. Artificial Neural Network.....	26
iii. Regression .....	28
3.3 <i>k</i> -fold Cross Validation.....	30
3.4 Performance Evaluation Techniques .....	31
IV. DESCRIPTIVE PROFILING.....	32

4.1 Variable analysis.....	32
4.2 Hypothesis Testing.....	62
V. PREDICTIVE MODELING .....	79
5.1 Balanced Data With IBM SPSS Modeler .....	79
i. Neural Network: Multilayer Perceptron (MLP) .....	79
ii. Neural Network: Radial Basis Function (RBF) .....	82
iii. Decision Trees .....	85
5.2 Unbalanced Data With IBM SPSS Modeler.....	86
i. Neural Network: Multilayer Perceptron (MLP) .....	86
ii. Neural Network: Radial Basis Function (RBF) .....	88
iii. Decision Trees .....	90
5.3 Balanced Data With SAS.....	92
i. Neural Network: Multilayer Perceptron (MLP) .....	92
ii. Neural Network: Radial Basis Function (RBF) .....	95
iii. Decision Trees .....	99
5.4 Unbalanced Data With SAS.....	102
i. Neural Network: Multilayer Perceptron (MLP) .....	102
ii. Neural Network: Radial Basis Function (RBF) .....	104
iii. Decision Trees .....	107
VI. CONCLUSION.....	112
REFERENCES .....	117
APPENDICES .....	120

## LIST OF TABLES

Table	Page
1. Cancer types with estimated cases .....	3
2. Types of Malignant Neoplasm of Ovaries .....	9
3. Types of Discharged Classifiers .....	17
4. Types of Filtered Discharged Classifiers .....	17
5. Frequency distribution for different Malignant Neoplasm of Ovaries .....	33
6. Variable analysis: Age of a patient .....	33
7. Variable analysis: Marital Status of a patient .....	35
8. Variable analysis: Patient type .....	41
9. Frequency distribution for different Admission sources .....	43
10. Frequency distribution for ovarian cancer patients with present on admit .....	45
11. List of Census Division .....	54
12. Variable analysis: Payer type .....	59
13. Variable analysis: Total Charges billed .....	61
14. Variable analysis: Length of stay of a patient .....	61
15. Statistics for table for Survival vs type of healthcare care setting .....	62
16. Cross Tabulation: Survival vs type of healthcare care setting .....	63
17. Cross Tabulation: Survival vs Bed size category of hospital .....	64

18. Statistics for table for Survival vs Bed size category of hospital .....	65
19. Cross Tabulation: Survival vs Census Region .....	66
20. Statistics for table for Survival vs Census Region.....	66
21. Statistics for table for Survival vs Race.....	67
22. Cross Tabulation: Survival vs Race.....	68
23. Statistics for table for Survival vs Martial Status .....	69
24. Cross Tabulation: Survival vs Martial Status .....	70
25. Statistics for table for Survival vs Payer type.....	71
26. Cross Tabulation: Survival vs Payer type.....	72
27. Statistics for table for Survival vs Age of patient.....	74
28. Cross Tabulation: Survival vs Age of patient.....	75
29. Statistics for table for Survival vs Length of stay of a patient.....	76
30. Cross Tabulation: Survival vs Length of stay of a patient.....	77
31. Statistics for table for Survival vs Total charges billed.....	78
32. Cross Tabulation: Survival vs Total charges billed.....	78
33. NN- MLP balanced data using IBM SPSS .....	80
34. Performance evaluation NN- MLP balanced data using IBM SPSS .....	82
35. NN- RBF balanced data using IBM SPSS.....	83
36. Performance evaluation NN- RBF balanced data using IBM SPSS.....	84
37. Performance evaluation decision tree balanced data using IBM SPSS .....	86
38. NN- MLP unbalanced data using IBM SPSS .....	87
39. Performance evaluation NN- MLP unbalanced data using IBM SPSS .....	88
40. NN- RBF unbalanced data using IBM SPSS.....	89

41. Performance evaluation NN- RBF unbalanced data using IBM SPSS .....	90
42. Performance evaluation decision tree unbalanced data using IBM SPSS .....	91
43. Predictor importance NN- MLP balanced data using SAS .....	94
44. Statistics for NN- MLP balanced data using SAS .....	94
45. Performance evaluation NN- MLP balanced data using SAS .....	95
46. Predictor importance NN- RBF balanced data using SAS .....	98
47. Statistics for NN- RBF balanced data using SAS .....	98
48. Performance evaluation NN- RBF balanced data using SAS .....	99
49. Predictor importance decision tree balanced data using SAS .....	100
50. Statistics for decision tree balanced data using SAS .....	101
51. Performance evaluation decision tree balanced data using SAS .....	101
52. Predictor importance NN- MLP unbalanced data using SAS .....	103
53. Statistics for NN- MLP unbalanced data using SAS .....	104
54. Performance evaluation NN- MLP unbalanced data using SAS .....	104
55. Predictor importance NN- RBF unbalanced data using SAS .....	106
56. Statistics for NN- RBF unbalanced data using SAS .....	107
57. Performance evaluation NN- RBF unbalanced data using SAS .....	107
58. Predictor importance decision tree unbalanced data using SAS .....	108
59. Statistics for decision tree unbalanced data using SAS .....	109
60. Performance evaluation decision tree unbalanced data using SAS .....	109
61. Overall performance evaluation balanced data .....	110
62. Overall performance evaluation unbalanced data .....	111

## LIST OF FIGURES

Figure	Page
1. Ovarian Cancer .....	4
2. CRISP- DM process.....	19
3. Process flow of data extraction .....	21
4. Graphical representation of NN with multilayers.....	27
5. Graphical representation of NN with activation function.....	28
6. Distribution Chart of age in years .....	34
7. Survival rate based on Age .....	34
8. Bar chart of Relative Percentage of age .....	35
9. Pie chart of marital status for ovarian cancer patients.....	36
10. Survival rate based on marital status .....	36
11. Bar chart of Relative Percentage marital status .....	37
12. Bar chart of Marital Status vs Survived =1.....	37
13. Bar chart of race of ovarian cancer patients.....	39
14. Survival rate based on race .....	39
15. Bar chart of relative percentage race .....	40
16. Bar chart of race vs Survived =1.....	40
17. Survival rate based on patient type .....	41
18. Bar chart of Relative Percentage of patient type .....	42

19. Bar chart of patient type vs Survived =1 .....	42
20. Survival rate based on Admission source .....	43
21. Bar chart of Relative Percentage of Admission source .....	44
22. Bar chart of Admission source vs Survived =1 .....	44
23. Pie Chart of Admission type of Ovarian Cancer patients .....	45
24. Survival rate based on Admission type.....	46
25. Bar chart of Relative Percentage of Admission type .....	46
26. Bar chart of Admission type vs Survived =1 .....	47
27. Pie chart of care-setting type for ovarian cancer patients .....	47
28. Survival rate based on care-setting type .....	48
29. Bar chart of Relative Percentage of care-setting.....	48
30. Pie chart of care-setting vs Survived =1 .....	49
31. Pie chart of urban vs rural status for ovarian cancer patients .....	49
32. Survival rate based on urban vs rural status.....	50
33. Bar chart of Relative Percentage of urban vs rural status .....	50
34. Bar chart of urban vs rural status vs Survived =1 .....	51
35. Pie chart of census region for ovarian cancer patients.....	52
36. Survival rate based on census region .....	52
37. Bar chart of Relative Percentage of census region .....	53
38. Bar chart of census region vs Survived =1 .....	53
39. Pie chart of census division for ovarian cancer patients .....	54
40. Survival rate based on census division .....	55
41. Bar chart of Relative Percentage of census division.....	55

42. Pie chart of census division vs Survived =1 .....	56
43. Pie chart of bed size range of hospital .....	57
44. Survival rate based on bed size range .....	57
45. Bar chart of Relative Percentage of bed size range .....	58
46. Pie chart of bed size range vs Survived =1 .....	58
47. Survival rate based on Payer type .....	59
48. Bar chart of Relative Percentage of Payer type .....	60
49. Pie chart of Payer type vs Survived =1 .....	61
50. Predictor importance NN- MLP balanced data using IBM SPSS.....	81
51. Predictor importance NN- RBF balanced data using IBM SPSS .....	84
52. Predictor importance decision trees balanced data using IBM SPSS .....	85
53. Predictor importance NN- MLP unbalanced data using IBM SPSS.....	87
54. Predictor importance NN- RBF unbalanced data using IBM SPSS .....	90
55. Predictor importance decision trees unbalanced data using IBM SPSS .....	91
56. Decision trees NN-MLP balanced data using SAS.....	93
57. Decision trees NN-RBF balanced data using SAS .....	97
58. Decision trees for balanced data using SAS .....	100
59. Decision trees NN-MLP unbalanced data using SAS.....	103
60. Decision trees NN-RBF unbalanced data using SAS .....	106
61. Decision trees for unbalanced data using SAS .....	108

## CHAPTER I

### INTRODUCTION

Over a few decades, the healthcare domain has been using IT for varied purposes such as storing patients visit details, cost, insurance details and more. Healthcare data are massive, data warehousing, knowledge management techniques can contribute to decision support systems in healthcare. The task of data collection and storage has improved extensively not just in terms of data collection, but also in its volume. Analyzing such an enormous amount of data would require specialized tools to analyze it, as manual data analysis would be tedious for such voluminous data. Medical informatics incorporate such needs by the use of statistical pattern recognition, machine learning, and visualization tools that would support the analysis of the data that are encoded in the given data, using a new interdisciplinary field of knowledge discovery in databases (KDD) (Frawley, Piatetsky-Shapiro, & Matheus, 1992). The method of data collection, analysis and formulation of knowledge out of these recognized patterns and visualization is referred to as data mining (Cios & Moore, 2002). Knowledge discovery in databases one of the main uses of data mining. Data mining tends to work well with such massive data. According to Frawley et al., 1992 “Data mining is the non-trivial extraction of implicit previously unknown and potentially useful information about data

According to Coorevits et al., 2013, the clinical research supported by electronic health records (EHR) is the upcoming new era. Since, millions of patients undergo treatments, equivalent or surplus amount of data is generated for these patients. Using these EHR specific knowledge and clinically actionable analysis can be generated. Various statistical, machine learning and computational techniques can be used to generate this specific knowledge, pattern recognition, clustering, generating models based on these EHR.

One of the challenging tasks in data mining is to use data mining to foresee the outcomes of a particular disease. One of those outcomes is survival. Survival analysis in the medical prognosis involves utilizing a set of parameters to predict the patients' health. The historic data of these patients is used to predict survivability.

According to the Centers for Diseases Control and Prevention, USA (2015) the number of deaths in 2014 were 2,596,993. Out of these 2,596,993 the top five death causing diseases were:

1. Heart disease: 611,105 deaths
2. Cancer: 584,881 deaths
3. Chronic lower respiratory diseases: 149,205 deaths
4. Accidents (unintentional injuries): 130,557 deaths
5. Stroke(cerebrovascular diseases): 128,978 deaths

Cancer is the second death causing disease and 1,658,370 new cases of cancer are estimated and out of these 589,430 people will die from the disease (National Cancer Institute 2015). Table 1 show the list of estimated new cases and estimated deaths due to different cancers.

<b>Cancer type</b>	<b>Estimated new cases</b>	<b>Estimated deaths</b>
Bladder	74,000	16,000
Breast (Female- Male)	231,840 - 2,350	40,290-440
Colon and Rectal (combined)	132,700	49,700
Endometrial (Ovarian)	54,870	10,170
Kidney (Renal Cell and Renal Pelvis)	61,560	14,080

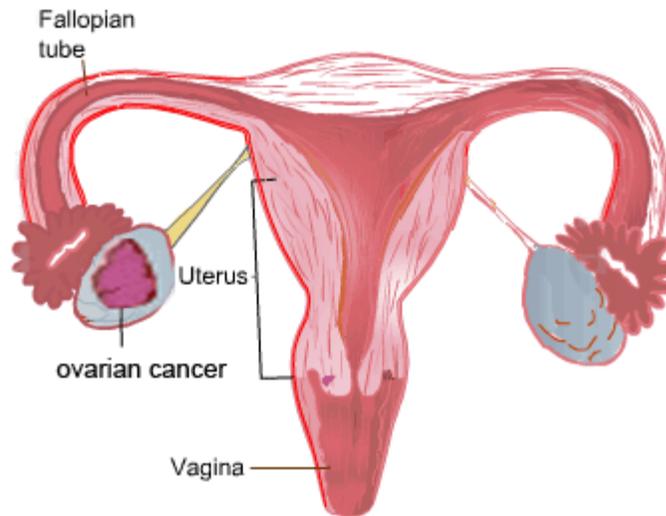
**Table.1** Cancer types with estimated cases.

This research focuses on predicting ovarian cancer survivability. The reasons that motivated this research on are:

1. The serious effects of ovarian cancer
2. Limited prior studies in the field
3. Potential of data mining technique and
4. A desire to further understand the nature of ovarian cancer

### **1.1 About Ovarian Cancer**

Ovarian cancer is a disease that is caused by malignant or cancerous cell found in the reproductive glands of women. According to American Cancer Society, 2015: “Ovarian cancer begins in the ovaries. Ovaries are reproductive glands found only in females (women). The ovaries produce eggs (ova) for reproduction. The eggs travel through the fallopian tubes into the uterus where the fertilized egg implants and develops into a fetus. The ovaries are also the main source of the female hormones estrogen and progesterone. One ovary is on each side of the uterus in the pelvis.” As shown in Fig.1.



**Fig.1** Ovarian Cancer. (<http://www.ovarydisease.com/p/ovarian-cancer.html>)

## 1.2 Signs and Symptoms

Ovarian cancer is difficult to detect during the early stages. In most cases the signs and symptoms are seen in advanced stages. The potential signs and symptoms may include:

- Abdominal bloating or swelling
- Pain in the belly or pelvis
- Frequent urination
- Loss of appetite or feeling full too soon

There are other symptoms which if they occur more than 12 times a month, may also be ovarian cancer symptoms:

- Fatigue
- Constipation
- Menstrual changes
- Back pain

### **1.3 Types of ovarian cancer**

Ovaries are made of three types of cell; each type cell can develop its own tumor. These tumors can be classified into benign (non-cancerous) and malignant (cancerous). The benign type of tumor never spreads beyond the ovaries. However, it can be treated by removing either the ovary or the part of the ovary which contains the tumor. The malignant type of tumor can spread to the other parts of the body and can be fatal. The three types of tumors are as follows:

#### **a) Epithelial tumors**

Epithelial tumor starts from the epithelial cells: that cover the outer surface of the ovaries. The epithelial tumor are further classified into three sub types: benign (non-cancerous), low malignant potential (low potential cancer) and malignant (high potential cancer). About 90% of ovarian cancers are epithelial tumors.

#### **b) Stromal tumors**

Stromal tumors develop from the connective tissue cells that produce female hormones. This is a very rare class of tumors; it accounts for about 1% of ovarian cancers.

#### **c) Germ cell tumors**

The germ cell tumors begin from the cells that produce eggs. According to the American Cancer Society (2015), this type of tumor accounts for less than 2% of ovarian cancers and may be benign, but some can be life threatening.

### **1.4 Staging of Cancer**

Based on clinical examination and the findings at laparotomy, an assessment is made for the disease. Staging is very important because based on the stage each cancer type will need different treatments. Staging for the cancer needs to be done accurately, because if not, then cancer that has

spread outside the marked stage of that particular stage could be missed. The method used for staging is the International Federation of Gynecology and Obstetrics (FIGO) system. Using this system, the tumor first needs to be classified based on the results of surgery. The extent of the primary tumor is coded as T. The letter N is used to represent absence or presence of metastasis to nearby lymph nodes. And the presence or absence of distant metastasis is represented by the letter M. Once the patient's tumor is classified based on these letters, this information is combined with a Roman numeral form of coding, called stage grouping, which goes from stage I (least advanced stage) to stage IV (most advanced stage (Society of Gynecologic Oncology, 2014):

- **Stage I** Growth limited to ovaries
- **Stage II** Growth involving one or both ovaries with pelvic extension.
- **Stage III** Growth involving one or both ovaries with intraperitoneal metastases outside the pelvis
- **Stage IV** Growth involving one or both ovaries with distant metastases

### **1.5 Causes and Risk Factors**

Risk factors provide information about person's chances of acquiring the disease. These risk factors don't tell us everything, it is difficult to say how much did a factor contribute towards the cancer. A few factors are (Centers for Disease Control and Prevention, 2015; MedicineNet.com, 2015):

- Age: The higher the age, the higher the risk of developing ovarian cancer. Half of all ovarian cancers are found in women who are 63 and older.
- Obesity: Women with a BMI (Body Mass Index) of at least 30 have a higher risk of developing ovarian cancer.

- Estrogen therapy and hormone therapy: Recent studies suggest that the risk of developing ovarian cancer increases when estrogens is used after menopause.
- Family history of ovarian cancer, breast cancer, or colorectal cancer: Family history of some other cancer such as breast cancer, or colorectal cancer increases the risk of ovarian cancer.
- Family cancer syndromes: 5-10 % of ovarian cancer are due to family cancer syndromes, where mutation in the genes BRCA1 and BRCA2 are responsible for most inherited ovarian cancer.
- Fertility drugs: Some researchers have found using fertility drug clomiphene citrate (Clomid®) for a period longer than one year increases the risk of ovarian cancer.

## 1.6 Diagnosis

Diagnosis of ovarian cancer is possible. Methods for diagnosis include imaging test like computed tomography (CT) scans, magnetic resonance imaging (MRI) scans and ultrasound studies can confirm whether a pelvic mass is present. Other tests like laparoscopy, colonoscopy, biopsy, and blood test can help if a woman shows symptoms of ovarian cancer. According to the American Cancer Society, 2015 about 20% of ovarian cancer are found at an early stage. There are a few ways to find ovarian cancer early(Cancer.Net, 2015; National Ovarian Cancer Coalition, 2015)

- Regular women's health exams: A regular health exam including the pelvic exam may help identify any cancerous tumor in the pelvic area.
- Visit the doctor if symptoms are seen: During the early stages, cancer doesn't cause many symptoms, but if symptoms are present over a longer period of time, like 12 times during the course of a month, gynecologists suggest seeking medical attention.

- Screening test for ovarian cancer: Screening tests help to find cancer even for people who don't show symptoms of ovarian cancer. There has been a lot of research so far to find the best way for screening ovarian cancer, but there is no single method that has been completely reliable. However, CA-125 blood test and transvaginal ultrasound (TVUS) are the most often used tests. TVUS can help find tumor in the ovary, but it can't show if the tumor is benign or malignant. When TVUS is used there can be cases where the tumor found might not be cancerous. CA-125 is used to test the protein level in the blood. The problem with using this test is there can be other conditions that can cause high levels of CA-125. Studies found that TVUS and CA-125 are used a lot for screening and testing, but it did not lower the number of deaths caused by ovarian cancer.

According to the American Cancer Society (2015), in the United States about 21,290 women will receive a diagnosis of ovarian cancer in 2015. About 14,180 U.S. women are estimated to die from ovarian cancer in 2015. This study deals with all types of Malignant Neoplasm of Ovary listed in Table 2. The diagnosis type for all of the listed malignant neoplasm of ovary is ICD9. According to Centers for Disease Control and Prevention, 2015 (ICD-9-CM) is:

“The International Classification of Diseases, Ninth Revision, Clinical Modification (ICD-9-CM) is based on the World Health Organization's Ninth Revision, International Classification of Diseases (ICD-9). ICD-9-CM is the official system of assigning codes to diagnoses and procedures associated with hospital utilization in the United States.”

<b>Diagnosis codes</b>	<b>Diagnosis description</b>
183	Malignant Neoplasm of Ovary and Other Uterine Adnexa
183.0	Malignant Neoplasm of Ovary
183.2	Malignant Neoplasm of Fallopian Tube
183.3	Malignant Neoplasm of Broad Ligament of Uterus
183.4	Malignant Neoplasm of Parametrium
183.5	Malignant Neoplasm of Round Ligament of Uterus
183.8	Malignant Neoplasm of Other Specified Sites of Uterine Adnexa
183.9	Malignant Neoplasm of Uterine Adnexa, Unspecified

**Table.2** Types of Malignant Neoplasm of Ovaries

### **1.7 Problem Statement**

The aim of this study is to investigate patterns or factors that influence the survival of ovarian cancer patients using predictive modeling techniques in data mining. Using predictive modeling the study aims at answering a few research questions like:

- Are there any particular sets of patterns that are important to the survival of the patient?
- Given a set of non-modifiable factors (race of the patient, admission type of the patient, length of stay and so on.), is it possible to predict whether or not an ovarian cancer patient would survive or not survive? Accepting model accuracy of more than 75%.
- Evaluating the predictive models using balanced and unbalance datasets. Thus, suggest which technique to use while using a real world unbalanced data.

### **1.8 Research Objectives**

Ovarian cancer has not been a great focus of research using data mining techniques. While other cancers like prostate cancer, breast cancer have already been studied, it is important to study ovarian cancer. In addition, as large number of studies have investigated the early detection of ovarian cancer and numerous studies have published that symptoms may not occur until the last

stage of ovarian cancer. As there are no diagnostic tools to detect ovarian cancer, symptoms remain the most important factor for its detection. The study aims at:

1. Providing a different approach to predict the survival of ovarian cancer patients using machine-learning techniques given a set of non-modifiable factors vary from race of the patient, admission type of the patient, length of stay and so on.
2. To approach this issue use descriptive profiling: finding patterns from the data.
3. Test various hypothesis on the data.
4. Using predictive modeling approach, which is also known as supervised prediction or supervised learning. This technique uses two sets of variables called *inputs* i.e. predictors, features, explanatory variables and other set called *targets* i.e. outcome, dependent variable. It can also be said that predictive modeling outputs are the predictions or guesses of the given set of inputs or predictor variables. To do so this research aims to take advantage of the available historic data of the patients by using three classification data mining techniques like Decision trees, Artificial Neural Network- Multilayer Perceptron (MLP) and Artificial Neural Network-Radial Basis Function (RBF).
5. In addition, find which machine-learning techniques performs better using balanced and unbalance datasets.
6. Accepting a model that predicts ovarian cancer survival with an accuracy of more than 80%

## CHAPTER II

### LITERATURE REVIEW

There are a large number of studies that are investigating the early detection of ovarian cancer. Numerous studies have published that symptoms may not occur until the last stage of ovarian cancer. As there are no diagnostic tools to detect ovarian cancer, symptoms remain the most important factor for its detection. However, two studies done in the United States and United Kingdom studying the screening of ovarian cancer found that the CA-125 protein level in blood detected more cancer. The results of this screening didn't lead to any better outcome than for those who weren't screened.

New ways of treatment including new testing methods are also being studied by researchers. A study carried on germline mutations in a gene on chromosome 17q known as BRCA1 show that BRCA1 genes are responsible for a large proportion of inherited breast cancer and ovarian cancer (Ford, Easton, Bishop, Narod, & Goldgar, 1994). A study on survival benefit showed that second-look laparotomy after completion of first line single agent cisplatin chemotherapy did not show any benefits for survival for patients with epithelial ovarian cancer(Luesley et al., 1988). One study on effective screening of ovarian cancer shows that if the cancer is screened before it

metastasizes, the detection of preclinical disease at an early stage would improve the overall survival (Bast Jr et al., 2002). A case based study on 150 patients studying the influence of fertility and oral contraceptives on the risk of ovarian cancer on patients under the age of 50, shows that women who had an immediate intolerance to oral contraceptive showed increased risk of ovarian cancer(Casagrande et al., 1979).

There are studies that focus only a certain age groups, such as one study on analysis of outcomes with neoadjuvant chemotherapy or primary cytoreduction that was performed on elderly ovarian cancer patients. The study found that the patient population aged 80+ didn't show any variation in complication rate for chemotherapy and surgical related complications when compared to those aged 65-79. This study concludes there didn't appear to be any impact of the choice of initial treatment on the survival when these decisions were used in a selective manner (McLean et al., 2010). Elderly patients with advanced ovarian cancer receive less frequent chemotherapy was the conclusion from Sundarajan et al.,2002 done on a patients over age 65 (Sundararajan, Hershman, Grann, Jacobson, & Neugut, 2002).

Further studies based on chemotherapy for elderly patients were carried out by Eisenhauer et al.,2007, who concluded that patients aged 65+ or 65 showed no differences on initial response, platinum resistance, PFS and OS as compared to other younger patients. It was also noted in the study that elderly women who can handle primary cytoreductive surgery should receive platinum-taxane chemotherapy along with it (Eisenhauer et al., 2007). Most studies did not find any variation in complications produced by chemotherapy amongst different age groups. While some studies did find a particular age group to handle a certain type of chemotherapy well. Because of such contradictory outcomes, it is difficult to estimate if a particular age group would perform better than the other.

Various studies have been done on ovarian cancer using data mining one of which is applying data mining criteria to investigate 52 proteins being good candidates as ovarian cancer biomarkers(Kuk et al., 2009). A study on chronic disease prognosis and diagnosis system uses case based reasoning and data mining. This results for this study talks about how in the knowledge creating phase data mining techniques, the decision tree induction algorithm and the case association are used to discover implicit results from the data. This study uses rule which are stored in rule base for the particular chronic diseases prognosis. Based on those rules probabilities for new cases are calculated. These new cases will then trigger the case based reasoning mechanism to support cases from the library for that particular chronic diseases prognosis. The most important contribution of this study is that it shows how helpful implicit rules are which are based on the technique of data mining process(Huang, Chen, & Lee, 2007).

Jeetha & Malathi, 2013 intended to observe performance of Artificial Neural Network (ANN) over genetic algorithm on diagnosis of ovarian cancer. The conclusion of this study was to use ANN along with genetic algorithm and propose a method for refinement and categorizing the ovarian cancer with kind, spreading, and normal tissue. Since studies have focused on how a genetic algorithm performs when compared to machine learning technique, it is interesting to learn how these techniques be used as an advantage to learn about survival of ovarian cancer patients using non-modifiable factors.

Kumar & Bishoni, 2013 studied Reptree, BRN network, and simple logistic for diagnosis system, while concentrating on non-modifiable risk factors and modifiable risk factors for breast cancer survivability. Non-modifiable factors such as age, gender, menstrual history, age at menarche and age at menopause and modifiable factors like BMI, age at first child birth, number of years of breast feeding, alcohol, diet and number of abortion. This study suggests that simple logistic can be used to obtain fast automatic diagnostic system for other diseases. Breast cancer survivability has had a lot of focus over the last few years. One of such studies is by Delen et al.,

2005, the method used by this study is to use three data mining algorithms decision trees and artificial neural networks along with statistical method logistic regression. To compare the performance of these models a 10 fold cross validation strategy is used. The study finds decision tree (C5) to be the best predictor on the holdout sample with an accuracy of 93.6%, while logistic regressions show the worst accuracy of 89.2%. Similar study on breast cancer survivability on SEER dataset investigates three data mining techniques: the Naïve Bayes, the back-propagated neural network, and the C4.5 decision tree algorithms. The results found C4.5 algorithm has a much better performance than the other two techniques(Bellaachia & Guven, 2006).

According to Ahemad & Shereen, 2012 while comparing single decision tree (SDT), boosted decision tree (BDT) and decision tree forest (DTF) for the detection of breast cancer during validation phase of analysis DTF achieved 97.51 %, which was superior to SDT (95.75 %) and BDT (97.07 %) classifiers. Another study compares seven common algorithms (Logistic Regression model, Artificial Neural Network (ANN), Naive Bayes, Bayes Net, Decision Trees with naive Bayes, Decision Trees (ID3) and Decision Trees (J48)) besides the most widely used statistical method (Logistic Regression model) to find an optimal model to predict breast cancer survival rate. The study finds Logistic Regression model to be the optimal model with the highest accuracy of 85.8±0.2%. A study on predicting coronary artery disease (CAD) where strategies like logistic regression (LR), classification and regression tree (CART), multi-layer perceptron (MLP), radial basis function (RBF), and self-organizing feature maps (SOFM) are compared in order to predict if the patient has CAD. The comparison is made on the ROC curve, Hierarchical Cluster Analysis (HCA), and Multidimensional Scaling (MDS). The results of this study suggest MLP to be the best technique with an area under the curve for ROC as 0.783(Kurt, Ture, & Kurum, 2008).

Data mining has also been used in studying the performance of classification techniques for predicating risk of hypertension compares three decision trees, four statistical algorithms, and two

neural networks. The study concludes that ANN- MLP and RBF procedures, performed better than other techniques in predicting hypertension.(Ture, Kurt, Turhan Kurum, & Ozdamar, 2005). Stark & Pfeiffer,1999 compared logistic regression (LR), classification technique ID3, C4.5, CHAID, and CART. These techniques were compared on veterinary epidemiology dataset, they found classification techniques are well-suited for exploratory data analysis. Another study by King, Feng, & Sutherland, 1995 compared machine learning techniques like CART, C4.5, NewID,AC2, ITrule, Cal5 and CN2, statistics Naïve Bayes, k-nearest neighbor, kernel density, linear discriminant, quadratic discriminant, LR, projection suit and Bayesian networks and neural networks back propogation and RBF. These techniques were compared on twelve datasets with respect to large world problem.

Ovarian cancer has not been a thorough matter of research using data mining techniques; while other cancers like prostate cancer, breast cancer have already been studied in depth. In addition, large number of studies have investigated the early detection of ovarian cancer and many of them have concluded that symptoms may not occur until the last stage. As there are no diagnostic tools to detect ovarian cancer, symptoms remain the most important factor. That is why this study uses non-modifiable factors to predict ovarian cancer survivability by using similar comparative strategies on a real world dataset.

## CHAPTER III

### METHODOLOGY

#### **3.1 Introduction**

Data driven research in cancer is now becoming more prominent and useful, since most of the cancer research is usually clinical or biological in nature. The term survival analysis evolved from the initial studies, where the interest was death. Survival analysis in the medical prognosis field uses historic data to predict survivability. It can also be said that medical prognosis involves the use of prediction models where a patient's information of the disease is used to estimate his/her health. "Survival" as many dictionaries describe it is a state or fact of continuing to live or exist. Survival analysis also has been studied using events like death, recovery, relapse, and length of time an individual is in the hospital. Many researchers define a survival period as a 10 year duration, but over the years the definition has improved.

For defining "Survival" for the purpose of this study Table.3 has been used (shows the list of "discharge classifier") and for "Non-Survival" or "Expired" Table.4 has been used (shows the list of "Filtered Discharged Classifiers") which are derived from the dataset provided by *CHSI Cerner Health Facts*®. The discharge classifiers are indicator of the state of the patient's health

status (Table.3) also lists the indicators of their health after they are discharged home “Expired at home. Medicaid only, hospice” (Table.4).

No.	DISCHARGED CLASSIFIER
1	Discharged to home
2	Discharged/transferred to another short term hospital
3	Discharged/transferred to SNF
4	Discharged/transferred to ICF
5	Discharged/transferred to another type of inpatient care institution
6	Discharged/transferred to home with home health service
7	Discharged/transferred to home under care of Home IV provider
8	Discharged/transferred within this institution to Medicare approved swing bed
9	Discharged/transferred/referred another institution for outpatient services
10	Discharged/transferred/referred to this institution for outpatient services
11	Discharged/transferred to another rehab fac including rehab units of a hospital
12	Discharged/transferred to a long term care hospital.
13	Discharged/transferred to a nursing facility certified under Medicaid but not certified under Medicare.
14	Discharged/transferred to another Type of Health Care Institution not Defined Elsewhere
15	Discharged/transferred to a federal health care facility.
16	Discharged/transferred/referred to a psychiatric hospital of psychiatric distinct part unit of a hospital
17	Discharged/transferred to a Critical Access Hospital (CAH).
18	Discharged/Transferred to a designated cancer center or childrens hospital
19	Discharged for Other Reasons
20	Discharged to Care of Family/Friend(s)
21	Discharged to Care of Paid Caregiver
22	Discharged to Court/ Law Enforcement/Jail
23	Discharged to Other Facility per Legal Guidelines

**Table.3** Types of Discharged Classifiers

To define non-survival the dataset uses classifiers as listed in the table. All the different types of “expired” classifiers were later coded as “expired” for analysis (Table 4).

No.	DISCHARGED CLASSIFIER
1	Expired
2	Expired at home. Medicaid only, hospice.
3	Expired in a medical facility. Medicaid only, hospice.
4	Expired, place unknown. Medicaid only, hospice.

**Table.4** Filtered Discharged Classifiers

Cerner data warehouse was established with Health Insurance Portability and Accountability Act (HIPAA)–compliant operating policies and procedures using statistical methods for de-identification of clinical and financial information.

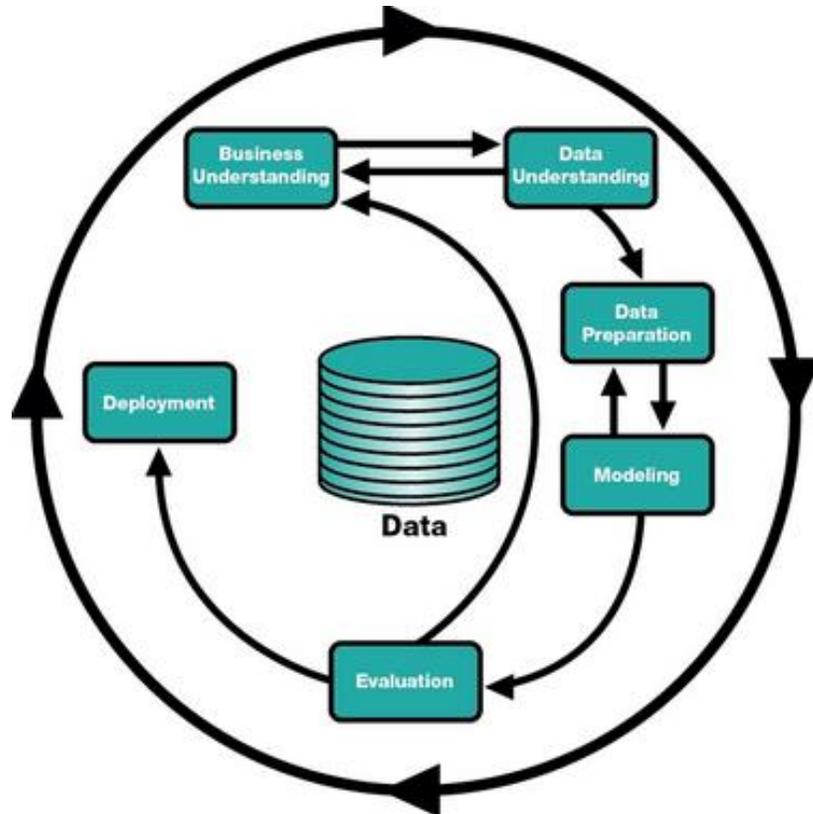
“The HIPAA Privacy Rule provides federal protections for individually identifiable health information held by covered entities and their business associates and gives patients an array of rights with respect to that information. At the same time, the Privacy Rule is balanced so that it permits the disclosure of health information needed for patient care and other important purposes.” (U.S. Department of Health & Human Services)

In simpler terms, once a patient are discharged home, HIPAA prevents the hospital from using patient’s information like address, phone number. Hence, it is difficult to keep track of a patient’s health once they have been discharged to home. Therefore, this is a limitation of this study that once a patient was been discharged to home there is a possibility that families do not report the status of patient’s health. Thus, such cases patient are been considered to survive. Accounting all of these factors and considering the in-depth status of patients while in hospital for the purpose of this using dataset provided by *CHSI Cerner Health Facts*® the “survival” of ovarian cancer patients has been defined as using the classifier “discharged home” as survived. The “discharged home” classifier chosen to mark a completely recovered patient because all other classifiers point to a patient who has not fully recovered (Table.3). While all the different types of “expired” classifiers were later coded as “expired” for analysis (Table.4).

### **3.2 CRISP- DM Approach**

In order to proceed with analysis of the data it is very important to have a data mining process. There are various data mining methodologies; these methodologies attempt to shape the way a data analyst approaches the steps to perform the data mining tasks. The two major methodologies that are most used are SEMMA and CRISP- DM. SEMMA, which stands for Sample, Explore,

Modify, Model, Assess, has been developed by SAS Institute. On the other hand, Cross Industry Standard Process for Data Mining (CRISP- DM) was developed by software vendors and industry users of data mining. For this study, the data mining process that is being used is a Cross Industry Standard Process for Data Mining (CRISP- DM) approach. This approach consists of six phases (see Fig 3.1):



**Fig.3.1** CRISP-DM process (<http://www.sv-europe.com/crisp-dm-methodology/>)

### 3.2.1. Business Understanding

This is the initial phase of the process where the business question or the question of interest is discovered. In this study the research question is, what factors contribute to survivability of an

ovarian cancer patient? Also, are there any patterns a patient needs to pay attention to in terms of getting early cancer treatment?

### **3.2.2. Data Understanding**

The given data is examined for the appropriate data type and knowledge is acquired about the various tables and variables available in the dataset.

### **3.2.3. Data Preparation**

To prepare the data for analysis the following sub steps were used:

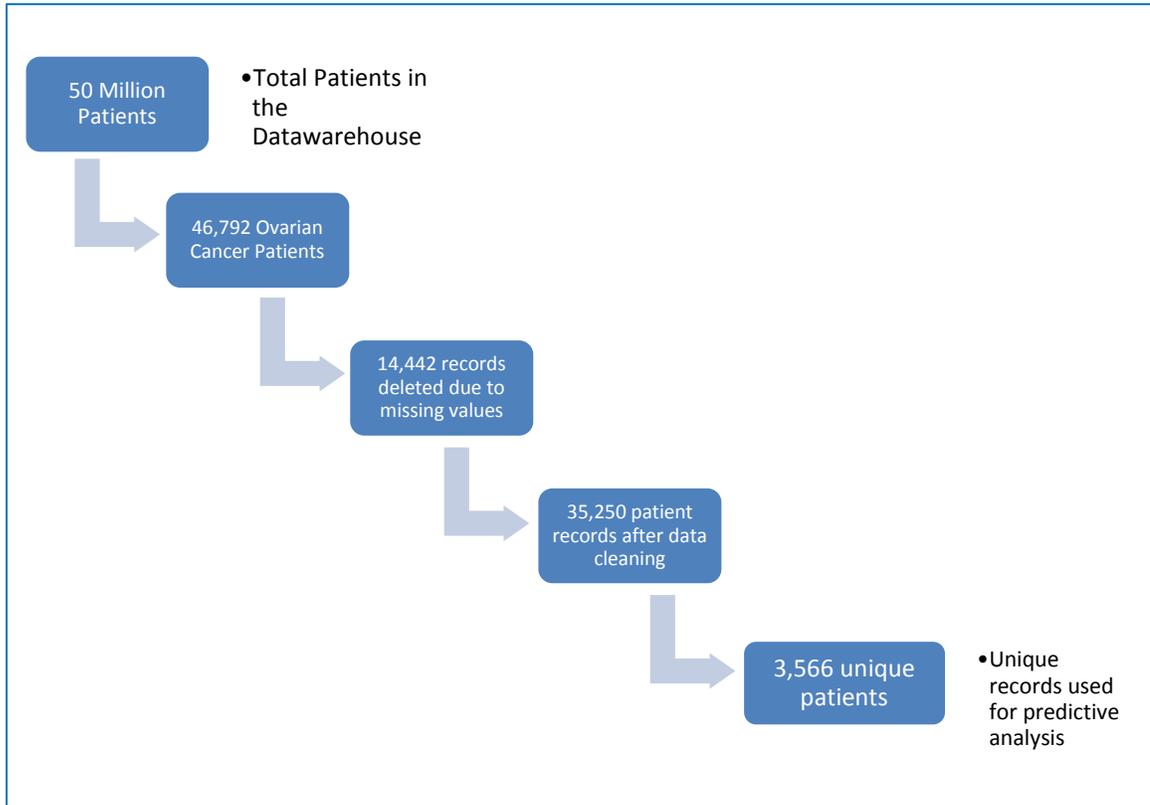
#### **a) Data Access**

The study uses the data from the Center for Health System Innovation (CHSI) provided by Cerner Health Facts. The Cerner Health Facts dataset is the largest relational database on healthcare. This database is a comprehensive source of de-identified, real-world, HIPAA-compliant data.

The database consists of more than 50 million unique patients, more than 2.4 billion laboratory results, more than 84 million acute admissions, emergency and ambulatory visits, more than 14 years of detailed pharmacy, laboratory, billing and registration data and more than 295 million orders for nearly 4,500 drugs by name and brand.

The data for ovarian cancer patients is extracted using the SQL Server Management Studio 2012. SQL is a standard language for accessing and manipulating databases. Since, *Cerner Health Facts*® is a Relational Database Management System the basis to use it is SQL. The flow chart Fig.3.2 shows the steps followed in the extraction process. There were more than 50 million unique patients, more than 14 years of detailed pharmacy, laboratory, billing and registration data and more than 295 million orders of drugs by name and brand. These were filtered and 46,792 ovarian cancer patients were considered. Later, these were brought down to 32,350 patients'

records which had the desired information related to survivability of these patients. Out of these 3,566 were unique patients, are useful for analysis of this study. These records were obtained using SQL and SAS software when required. Clauses, expressions, statements etc. were widely used in the data extraction process from SQL server.



**Fig.3.2** Process flow of data extraction

### **b) Data Consolidation**

The data was acquired in a text (.txt) file and was then brought into SAS enterprise Guide (6.1) this text file was converted into a .sas7data file. There were no major issues while consolidating the data.

### **c) Data Cleaning**

The raw text file consisted of 80 variables. The dataset had more than one Encounter\_ID, Hospital\_ID, Patient\_ID, Dischg\_disp\_ID, Procedure\_ID etc. so the duplicate columns were

deleted and only one of each column was kept. Other variables like `age_in_months`, `age_in_weeks`, and `age_in_days` etc. were dropped as all the records had all “0” values, this is observed using histograms and analyzing the mean, minimum and maximum. Since the data set also consisted of the variable `age_in_years` the other age related variables didn’t serve any purpose. Further, initial data exploration was conducted on each variable through descriptive statistics. For interval variables, the mean, minimum, maximum, and missing values were studied. Histograms were used to analyze the distribution of these interval variables. To check if there were any outliers box and whisker plots were used for interval variables. Very few records consisted of missing values. Once all the duplicate columns were removed and outliers were analyzed another step was essential to use the data for analysis. Since, each record in the data set captures the encounters of the patients visit to the facility. Which implies there can be multiple encounters for a particular patient.

As this study focuses on the survivability it was important to have single encounters of these patients. To do so the records were aggregated for the same person using the `patient_sk` variable that is unique for a particular patient. There are other data challenges that needs to be accounted for like the curse of dimensionality. The curse of dimensionality, an expression coined by mathematician Richard Bellman, refers to the exponential increase in data required to densely populate space as the dimension increases. In case of a large number of inputs the curse of dimensionality doesn’t fit the model so well. In this study, this issue was handled by reducing the total number of inputs from 80 to 45. The problem of redundancy occurs when the input doesn’t give any new information that has not already been explained. The dataset did have redundant variables for example: `dischg_disp_id` and `dischg_disp_code` these two variables explain the same information about the patients’ discharge status similarly other variables were discarded based on redundancy. After cleaning of data the variable list was brought down to 47 (Appendix).

#### **3.2.4. Modeling**

Data mining is also about model generation. This step involves using actual modeling techniques that can be used to address the research question. Data mining models can be generated for three types of tasks like: descriptive profiling, directed profiling and prediction. The descriptive models or descriptive profiling gives insight into what the data does. Outputs of these models and hypothesis testing generate graphs, charts and summary statistics of the data. Direct profiling is when the target and the inputs for the models are from the same time frame. Prediction is finding patterns in the data from one period that are capable of explaining outcomes in a later period. This study focuses on both Predictive modeling and descriptive profiling.

#### **3.2.5. Evaluation**

The task in this step is to evaluate the generated model and go back and forth to reach the desired outcome for the research question.

#### **3.2.6. Deployment**

The last step is report generation and summary of the project.

### **3.3 Descriptive Profiling**

To analyze data descriptively tools like histograms, bar charts, pie charts, and cross tabulation were used to see patterns or distribution of values in the fields. All of these techniques provide a better insight into how the variables are distributed and how diverse the data is. These techniques all help in drawing conclusions on what factors can be more significant, least significant and more. However, further analysis is always required to confirm these results.

### **3.4 Predictive Modeling**

Following the CRISP-DM approach helps a data analyst follow a sequence of steps that would help get standardized steps and results. There are a few data mining techniques where the model complexity increases as the number of terms increases, or the number of leaf nodes increases. The standard strategy for the model assessment of the generalized data is splitting. The dataset is split into three types of datasets: training, validation, and test. A portion of the raw dataset is used to build the models; this dataset is called the training dataset. The validation dataset is used to monitor and tune the performance of the built model. The test dataset is used to estimate the generalization of the generated models. For the purpose of this study, the data is partitioned into 66% Training dataset and 33% Validation dataset. Since, this partition is a popular partition ratio for most of the data mining applications.

#### **3.4.1 Balancing Technique**

After the data preparation process the dataset is found to have relatively fewer instances of patients who have “not survived” i.e. variable `survival_code= “0”` with 188 patients compared to the number of patients who have “survived” i.e. variable `survival_code= “1”` with 3,378 patients. So it can be said that the “survived” class label is the majority class and “not survived” is the minority class. The total number of records in the data is 3,566 records. A dataset is said to be imbalanced when the minority class contributes less than 35% (Li & Sun, 2012). Since, the minority numbers in the dataset used for this study are less than 35% this dataset can be said to be an imbalanced dataset. Many studies have found when an imbalanced data is used the results would tend to be biased towards the majority class, while not being so accurate about the minority class. There are three methods used to handle an imbalanced dataset: Random Under-Sampling (RUS), Random Over-Sampling (ROS), and Synthetic minority over sampling technique (SMOTE).

a) Random Under-Sampling (RUS)

In RUS strategy, cases from the majority classes are randomly removed without replacement so that there are approximately the same number of cases in the minority class as in the majority class.

b) Random Over-Sampling (ROS)

In this method, the minority classes are randomly added to the dataset without replacement until they are roughly of the same number as that of the majority class.

c) Synthetic minority over sampling technique (SMOTE)

In the SMOTE strategy, k nearest neighbors for each of the minority examples are found. Then these k nearest neighbors are randomly selected based on the over sampling rate. This generates a new synthetic case between the minority class and each of the found k nearest neighbors. The process of generating synthetic case is repeated until the number of cases are approximately the same in both classes (Chawla, Bowyer, Hall, & Kegelmeyer, 2002).

To balance the imbalanced data, for this study the method used is Random over-sampling method. The minority classes “not survived” i.e variable survival\_code= “0” are randomly added to the dataset without replacement until they are roughly of the same number as that of the majority class “survived” i.e variable survival\_code= “1”. After applying this method, the total number of records in the imbalanced data goes up to 6,574 records.

### **3.4.2 Predictive Modeling Classification Method:**

In this study, three classification methods are used: Decision trees, Artificial Neural Network-MLP and Artificial Neural Network- RBF since these are powerful and popular for both classification and prediction in the domain of predictive modeling. These techniques are used

along with  $k$ -fold cross validation technique is used to minimize the overfitting or bias in these predictive modeling techniques

### **i. Decision Trees**

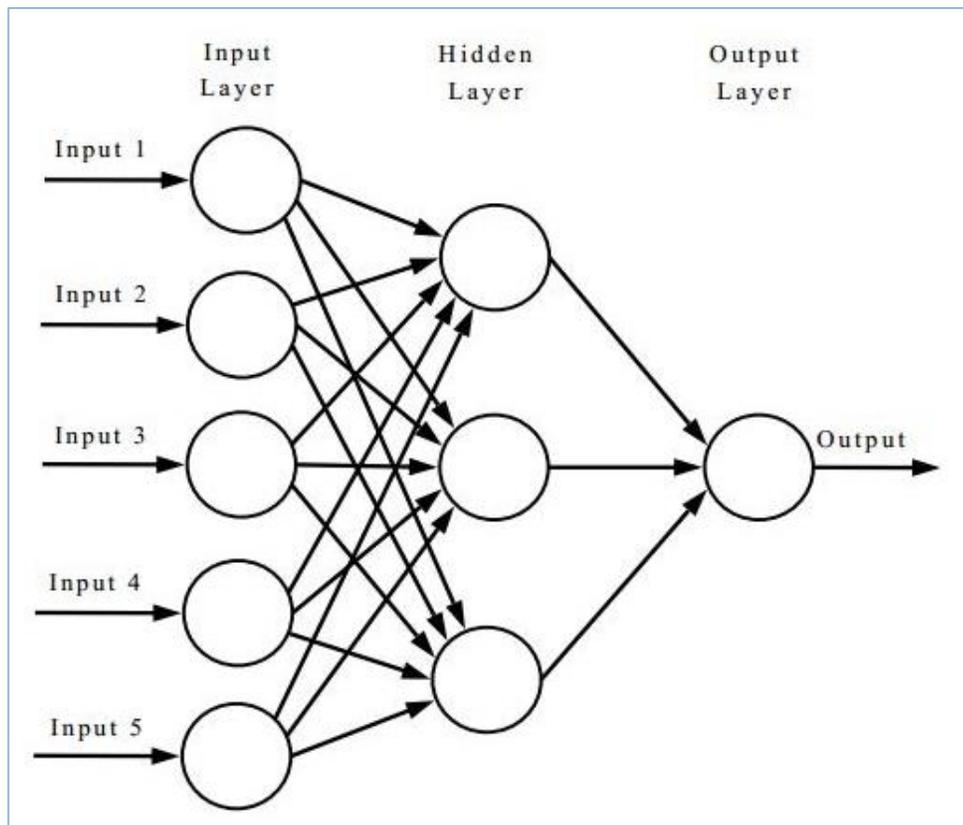
Decision trees can be defined as structures that can be used to develop a graphical user interface where large data is divided into successively smaller sets of records. Decision trees are governed by a set of rules that are used to produce successively smaller sets of records of the larger records. This technique is used to perform both classification and estimation tasks. The division of larger records (parent node) is called *splitting* and the successively smaller sets of records are called child nodes. The top of the tree is the root node, while the subsequent rules are interior nodes, and the end of the node or the node with only one connection are leaf nodes. Decision trees run on split search algorithms using different strategies to make splits. Since, the variable under observation is a binary variable “1” or “0”, we use “decision” to assess the subtree model. The advantages of using decision tree is they are not so complex structures. They are not sensitive to usual values (outliers) in the data and thus provide better performance. They reveal a lot about the data and require less data preparation.

### **ii. Artificial Neural Network**

Artificial Neural Network (ANN) are powerful tools and mathematical models used for pattern recognition and modeling. ANN behave in a similar fashion as that of the biological neurons found in nature. They are usually referred to as Neural Network and are used for classification, prediction, and clustering. NN consists for numerous formulations the study focuses on two formulation the multilayered perceptron (MLP) and radial basis function (RBF).

MLP is feed forward neural network (Fig.3.3) that is trained by the back propagation algorithm. It consists of several layers of computing units or neurons. Each unit or neuron belongs to the neural network that has a non-linear activation function. MLP contains one or more hidden layers

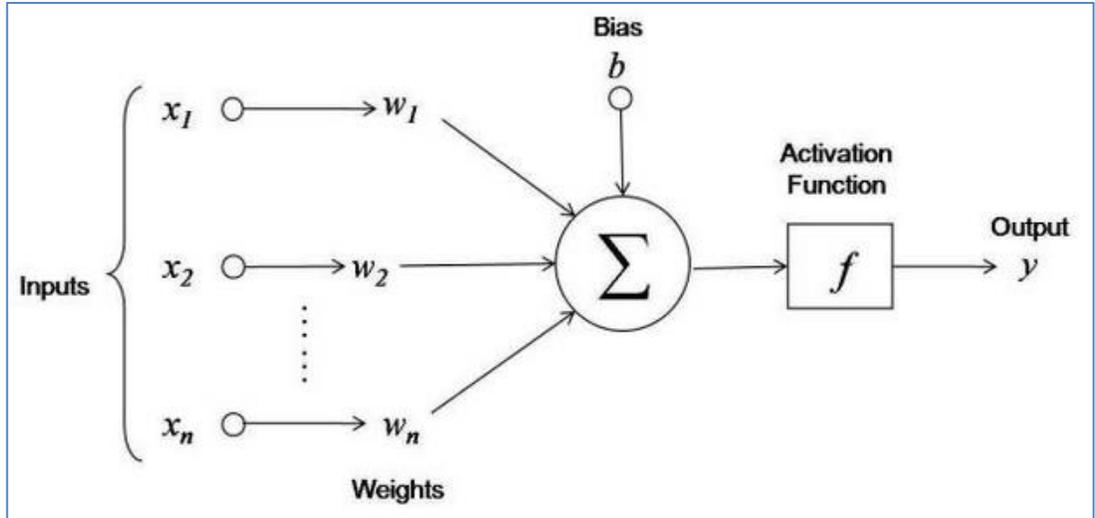
these layers are: input layer, hidden layer, and output layer. MLP networks are supervised techniques, they learn from the input data as to how to transform to desired response. Inputs for each layer are the previous layer inputs multiplied by the activation synaptic weights, summed and transformed by the activation function. These outputs are then sent to the next hidden layer.



**Fig.3.3** Graphical representation of NN with multilayers

RBF networks are non-linear and have a static bell shaped function. In this case, the activation of a hidden neuron or layer is based on the distance between the input vector and the center vector. It consists of input layer, hidden layer where there is a non-linear transformation from the input space to the hidden space and the output layer that generates the output for the network. In case of RBF divides the space into hyperspheres instead of using hyperplanes as in MLP. Thus, in this case the inputs are used to determine the basis function, which is followed by the finding weights

of the parameters found from the inputs. MLP minimizes the sum of squares of errors, which is also used in RBF.



**Fig.3.4** Graphical representation of NN with activation function

The advantages of using ANN is they usually perform better than the other models due to their complex structures. They are not sensitive to unusual values (outliers) in the data and thus provide better performance. The disadvantages of using ANN can be lack of clarity, because the network generated is not always easy to interpret. Since ANN are not capable of selecting inputs by themselves, they would use all the inputs present in the initial data. Sometimes due to large number of variables the ANN algorithm would not converge, so for the algorithm to converge ANN is used in combination with Logistic regression. Logistic regression in this case selects input variables for ANN algorithm and uses them for further analysis. Regression technique is explained as follows.

### **iii. Regression**

Regression models are the oldest prediction models: they create a specific association structure between inputs and targets. They use mathematical equations to predict cases using input

variables. There are two forms of regressions: linear and logistic regression. In a linear regression modeling approach, the target is predicted using simple combination of the input variables. Linear regression uses numeric input variables to predict the target numeric variable. They don't handle the cases with missing values as well as decision trees and neural networks. To use the cases which have missing values these missing values would need to be replaced or imputed. Imputation would use the mean, median or the mode to replace the missing values (blanks) with either of these values based on the variable type. Logistic regression are very much similar to linear regression. They use link function transformation for the target variable. As in linear regression, a linear combination of the inputs is used, but these inputs generate a logit score, the log of odds of primary outcome. The logistic function is inverse of the logit function. To predict the estimates, the logit equation for the target variables needs to be solved in this case. Regression also uses three sequential selection methods to select useful input variables. These three methods are: forward selection, backward selection and stepwise selection.

a) Forward Selection

The model starts with a base line model predicted by using overall average target values of all cases. The algorithm uses this base line model to search for new inputs for the models. A variable is added to the sequence only if the base line model shows a significant improvement in this complexity. Qualifier for the improvement is done on the bases of p-value, thus adding inputs increases the model's overall fit statistics. The algorithm terminates when there is no significant improvement in this complexity or the p-value. The p-value for this algorithm is predefined entry cutoff defined by the user based on their needs.

b) Backward Selection

In case of the backward selection the model starts with almost all possible input variables. The inputs are removed from the model and thus the complexity of the model decreases. Removing

inputs decreases the model's overall fit statistics. Qualifier for this method is done on the bases of p-value that is removing inputs with the highest p-value. The algorithm terminates when there is no significant improvement in this complexity or the p-value. The p-value for this algorithm is predefined stay cutoff defined by the user based on their needs.

c) *Stepwise Selection*

The model combines both the forward and the backward selection process. The model starts with a base line model predicted by using overall average target values of all cases, just like in the case of forward selection. A variable is added to the sequence only if the p-value is smallest and below the entry cutoff. Once this is done there is a reevaluation of the overall statistics of the model. Qualifier for the improvement is done on the bases of p-value, thus if the p-value of the added inputs increases the stay cutoff, the input is removed from the model. The variable once removed can be added and once the variable is added it can be removed too. The algorithm terminates when the variables that are added have greater p-values than predefined entry cutoff and also the p-value is below the predefined stay cutoff. .

### **3.3 k-Fold Cross Validation**

A k-fold cross validation technique is used to minimize the overfitting or bias in the predictive modeling techniques used in this study. k-Fold cross validation is used with decision tree and neural network with MLP and RBF to generate more flexible models to reduce overfitting or underfitting. The complete dataset R is randomly split into k-mutually exclusive subsets or folds of approximately equal size ( $R_1, R_2, R_3, \dots, R_k$ ). In classification model each of the k-subsets are trained and tested k times. One of the k-subsets is used as the test set and the other k-1 subsets are used as training dataset. Then the overall accuracy is calculated as the average of the k individual accuracy measures. All the machine learning techniques in this study use 10- fold cross validation.

### 3.4 Performance Evaluation Techniques

Evaluation of classification methods is one of the most important step in data mining. The most techniques are confusion matrix, learning curves and receiver operating curves (ROC). The confusion matrix shows the number of correct and incorrect prediction made by the prediction models. True positive (TP) are the case which are predicted correctly by the model, and actually the patient has survived. True negatives (TN) are those case that are predicted expired by the model, but actual is the patient has survived. False Positive are cases when model predicts the patient survived, but the patient actually expired (Type I error). False negative are cases when prediction of the model is the patient expired, but they actually have survived (Type I I error). Accuracy is the proportion of the total number of all the correct prediction of the model. It is calculated as the ratio between the total numbers of correctly classified cases to the overall number of cases under consideration.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

Sensitivity is the proportion of positive cases, which are correctly classified i.e. the percentage of patients who expired and are classified correctly as expired.

$$\text{Sensitivity} = \frac{TP}{TP + FN}$$

Specificity is the proportion of negative cases, which are correctly classified i.e. the percentage of patients who survived and are classified correctly as survived.

$$\text{Specificity} = \frac{TN}{TN + FP}$$

Where TP, TN, FP and FN denote true positives, true negatives, false positives and false negatives.

## CHAPTER IV

### DESCRIPTIVE PROFILING

The findings from descriptive analysis of the raw data set are discussed in this chapter. Section 1 discusses the variable analysis and section 2 discusses various hypothesis testing.

#### **4.1 Variable Analysis**

##### **Variable Analysis: Different Malignant Neoplasm of Ovaries**

Table 4.1 shows the distribution of ovarian cancer patients with different types of diagnosis codes. On observation, it was found there are 95% females with malignant neoplasm of ovary in this study. The remaining 5% of the females have been classified with others types of malignant neoplasms amongst which malignant neoplasm of fallopian tube is the second highest percent i.e. 3%. A much fewer number of females i.e. 0.03% are diagnosed with malignant neoplasm of ovary and other uterine adnexa.

diagnosis_code_desc	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Malignant Neoplasm of Broad Ligament of Uterus	3	0.08	3	0.08
Malignant Neoplasm of Fallopian Tube	105	2.94	108	3.03
Malignant Neoplasm of Other Specified Sites of Uterine Adnexa	26	0.73	134	3.76
Malignant Neoplasm of Ovary	3396	95.23	3530	98.99
Malignant Neoplasm of Ovary and Other Uterine Adnexa	1	0.03	3531	99.02
Malignant Neoplasm of Parametrium	29	0.81	3560	99.83
Malignant Neoplasm of Uterine Adnexa, Unspecified	6	0.17	3566	100.00

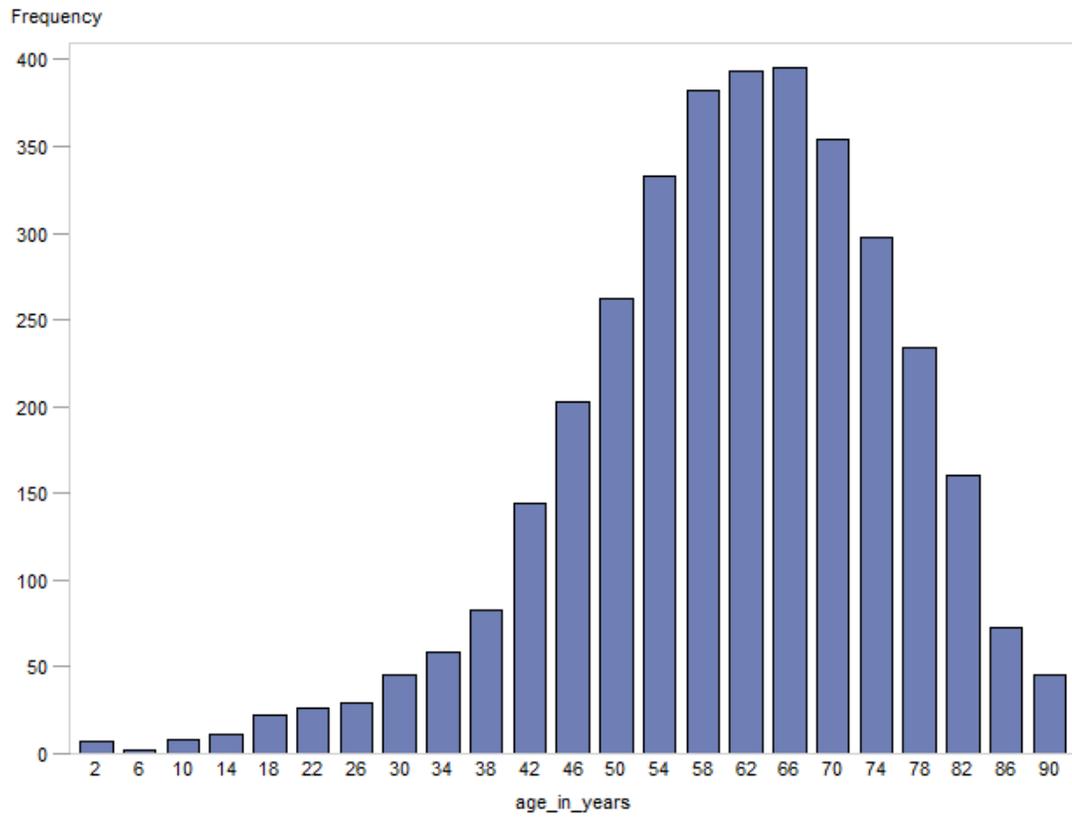
**Table.4.1** Frequency distribution for different Malignant Neoplasm of Ovaries

**Variable Analysis: Age**

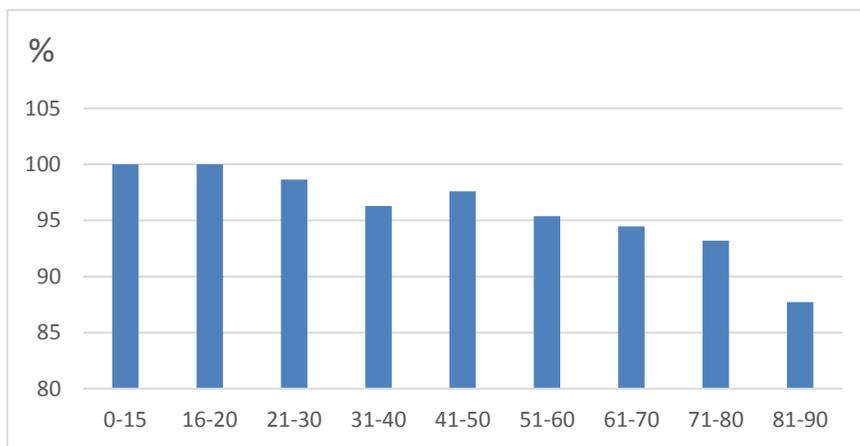
Table 4.2 shows the age distribution of the patients. It can be observed that the average age of the ovarian cancer patients in this study is 60 years old, while there seem to be a more number of 65 years old women in the case of this study. The dataset shows similar patterns observed by the SEER Cancer Statistics, which finds ovarian cancer rates highest in women aged 55-64 years. In this case, the mode is 65 years i.e. there are higher number of women who are aged 65 years. The age of the patients varies in the range from 0- 90 years. Also from the bar chart Fig.4.1, it can be said 63 year is the median age of this population.

Analysis Variable : age_in_years							
Mean	Std Dev	Variance	Minimum	Maximum	Mode	Range	N N Miss
60.0476725	14.7572779	217.7772498	0	90.0000000	65.0000000	90.0000000	3566 0

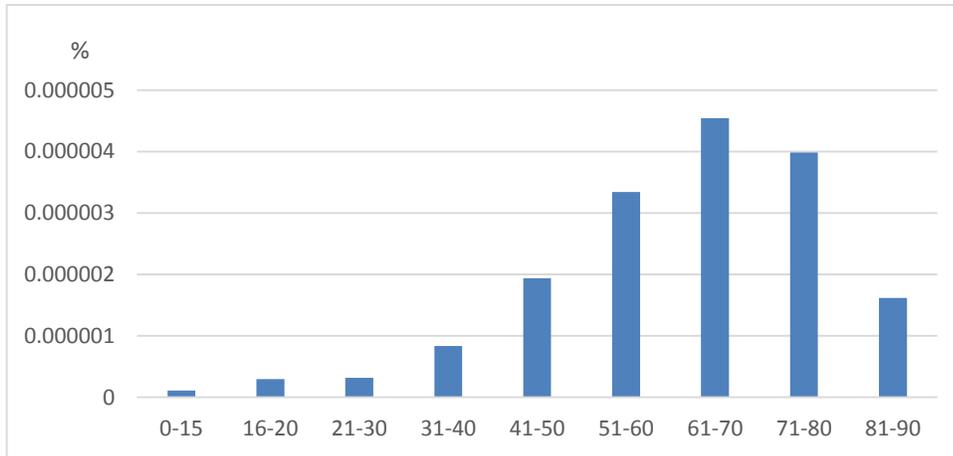
**Table.4.2** Variable analysis: Age of a patient.



**Fig.4.1** Distribution Chart of age in years



**Fig.4.2** Survival rate based on Age



**Fig.4.3** Bar Chart for Relative Percentages of age

Based on survival rate age group 16-20 years and 0-15 years shows better outcomes than other age groups (Fig.4.2). Figure 4.3 shows the relative percentages of age in years. The age group 61-70 years shows the maximum likelihood of having ovarian cancer, which is followed by age group 71-80. The relative percentage of different groups regarding a variable is calculated as:

$$\text{Number of ovarian cancer patients in a group} / \text{Number of all female patients (not just those with ovarian cancer) in the Cerner data set in that particular group.}$$

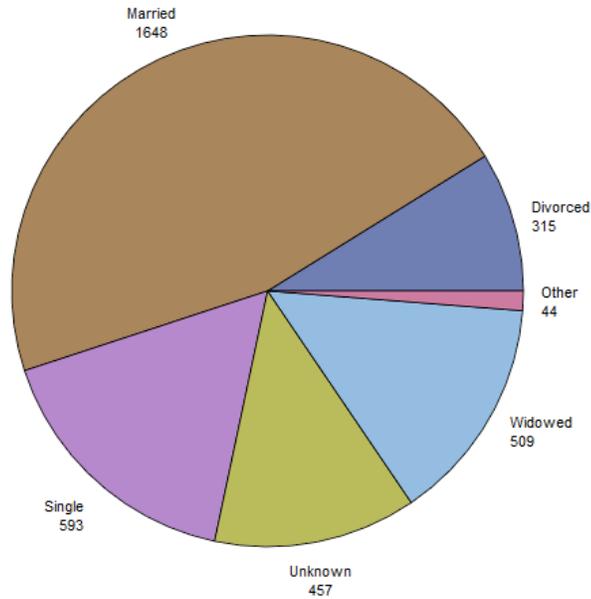
**Variable Analysis: Marital Status**

Table 4.3 clearly shows that “Married” women constitute 46% of the total ovarian cancer patients, while 17% “Single” ovarian cancer patients are single females.

Value	Proportion	%	Count
Married	46.21	46.21	1648
Unknown	12.82	12.82	457
Single	16.63	16.63	593
Widowed	14.27	14.27	509
Divorced	8.83	8.83	315
Legally Separated	1.15	1.15	41
Life Partner	0.08	0.08	3

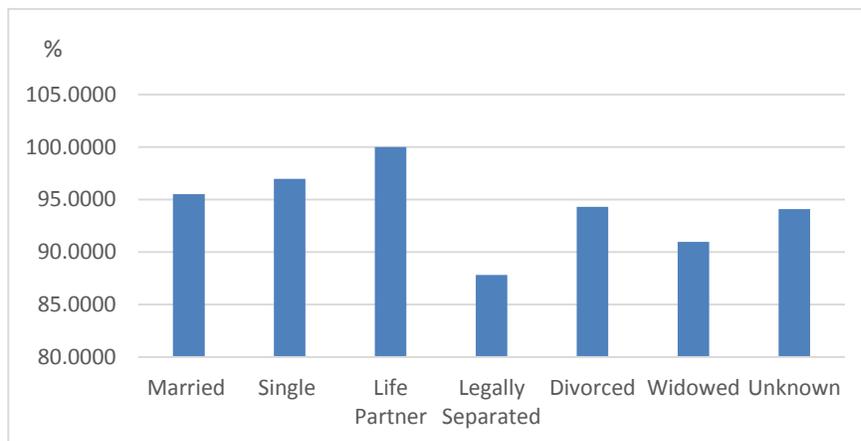
**Table.4.3** Variable analysis: Marital Status.

Pie chart Fig.4.4 shows a pictorial representation of the table above. Patients who have marital status life partner are 0.07% of all the types of marital status. As we see, the proportion of “Married” women is more than double that of “Single” women. Pie chart Fig.4.4 also confirms that there are 509 widowed women in this study, while women who live with life partners are the lowest with 3 in count.

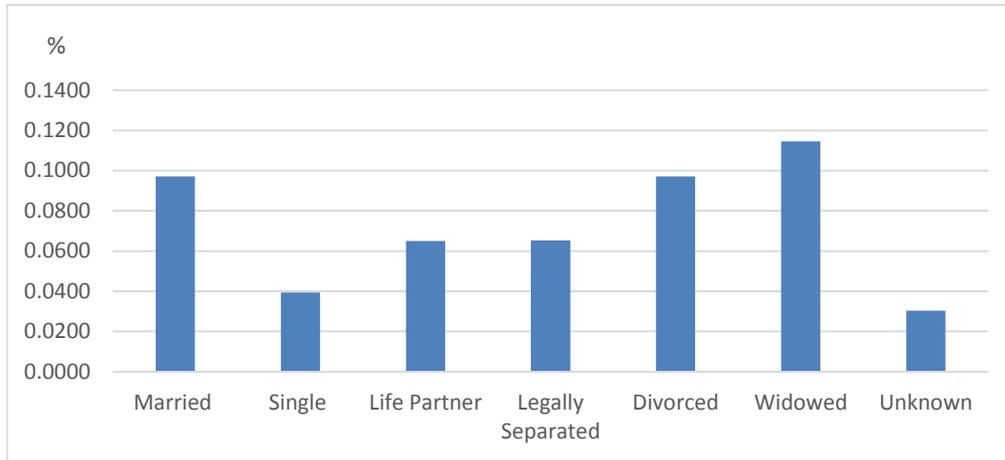


**Fig.4.4** Pie Chart of Marital Status

Figure 4.5 clearly points out the survival rate of ovarian cancer patients that “Married” women have 95% survival rate, while “Legally separated” women have 87% survival rate.

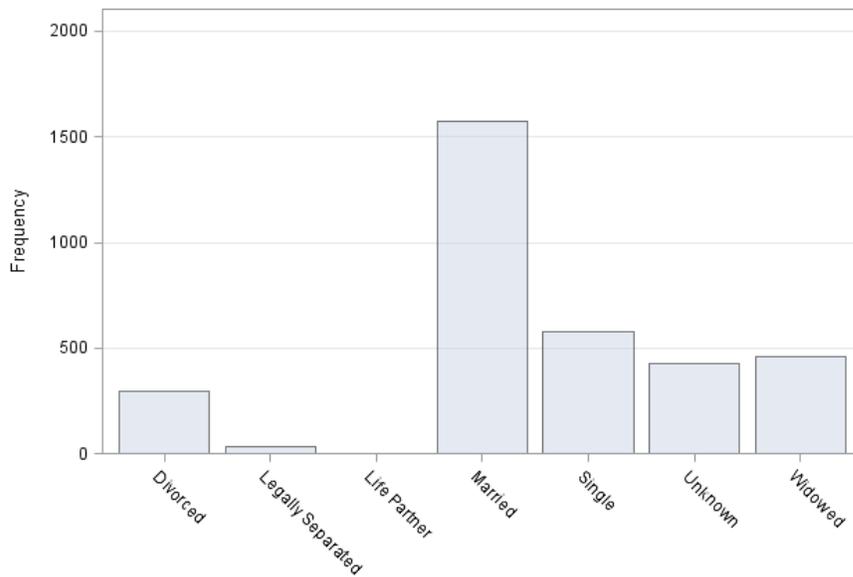


**Fig.4.5** Survival rate based on Marital Status



**Fig.4.6** Bar Chart of Relative Percentages of Marital Status

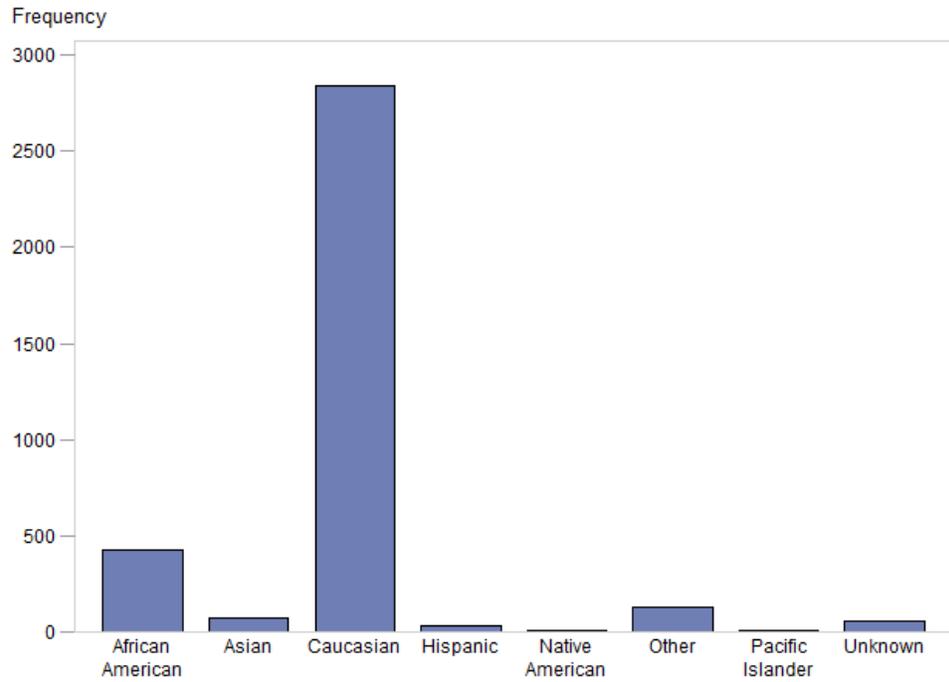
Bar chart Fig 4.6 shows a pictorial representation of the relative percentages of various marital status. The graph shows that “Widowed” women have the maximum likelihood of having ovarian cancer. Although an exact conclusion cannot be drawn on this graph alone. Figure 4.7 shows the distribution of survived patients with different marital status.



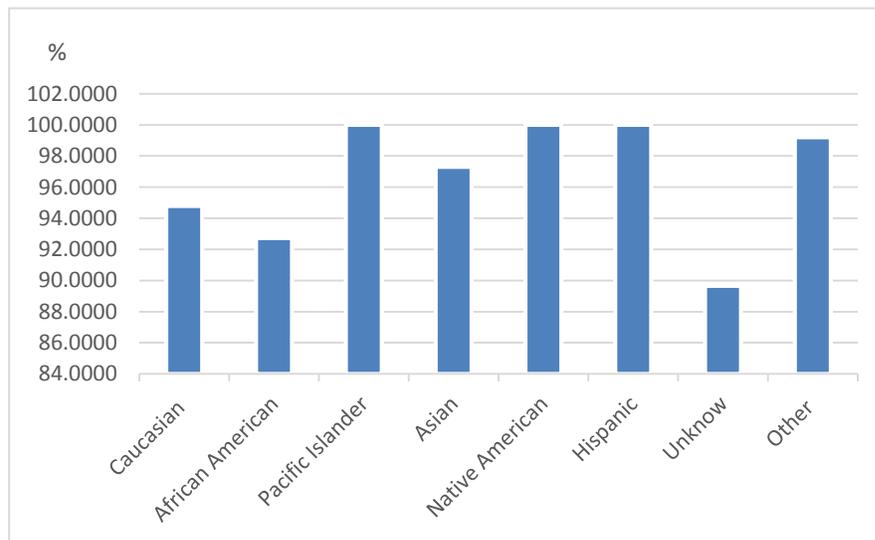
**Fig.4.7** Bar Chart of Marital Status vs Survived =1

### **Variable Analysis: Race**

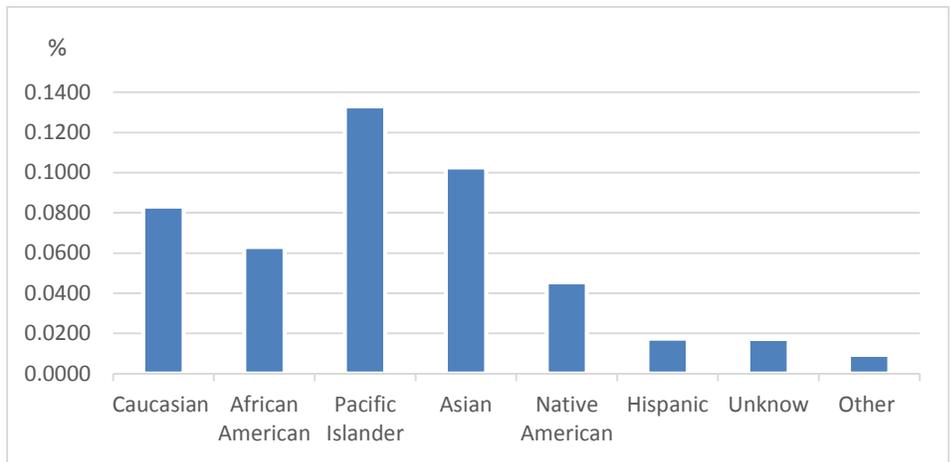
The next graph Fig.4.8 deals with the distribution of race in the dataset. It is observed from the graph that there are 2835 females who are “Caucasian”. “African American” women 462 in count are the second highest in this dataset. “Pacific Islander” women are the lowest in count. The data shows similar patterns as observed by SEER Cancer Statistics, which shows 13.0 new cases per 100,000 people are “White” and “Black” are 9.8 of new case per 100,000 people by race/ethnicity. Figure 4.10 shows relative percentages in this study, since this study consists of only Cerner’s datasets small portion of patients the percentages are relatively small. Also, if the females are healthy they would not be in the dataset. Figure 4.9 shows the survival rate of ovarian cancer patients based on race for this dataset. “Caucasian” women show a survival rate of 94%, while “African American” show 92% survival rate.



**Fig.4.8** Bar Chart of Race

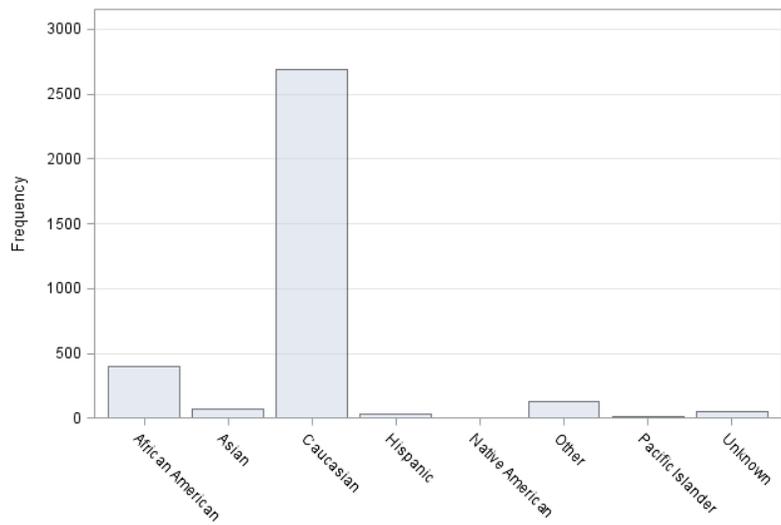


**Fig.4.9** Survival rate based on Race



**Fig.4.10** Bar Chart of Relative Percentages of Race

Figure 4.10 shows the relative percentage distribution of patients with different races. It is interesting to note that “Pacific Islander” and “Asian” women have relatively high likelihood of having ovarian cancer compared to women of “Caucasian” and other ethnicities. Figure 4.11 shows the distribution of survived patients with different race.



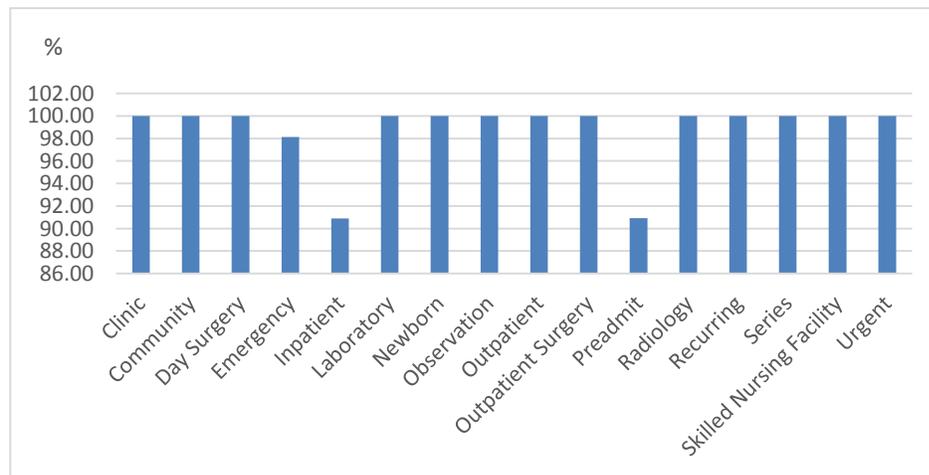
**Fig.4.11** Bar Chart of Race vs Survived =1

**Variable Analysis: Patient Type**

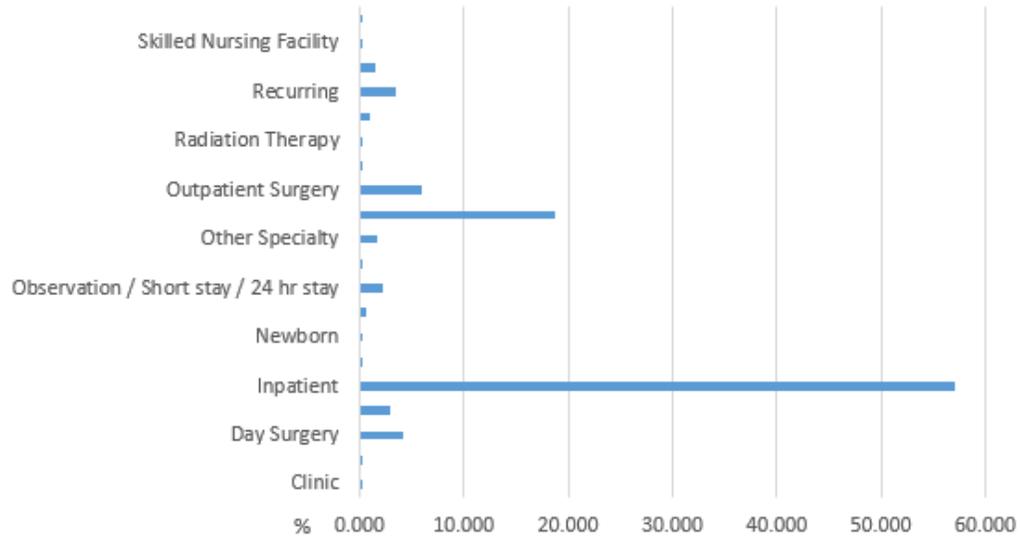
The data also captures the information whether the encounter (visit) was an inpatient, emergency room or outpatient visit. Table 4.4 shows the distribution of these patient visits. Most of the patient visits are “Inpatient” visits i.e. 2033 patients out of 3566, which is 57.01 % of the total patients. The “Emergency” visits are 3% of total patient visits. Figure 4.12 shows “Inpatients” have a survival rate of 91% and “Preadmit” shows survival rate of 90%.

PATIENT_TYPE_DESC	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Clinic	8	0.22	8	0.22
Community	3	0.08	11	0.31
Day Surgery	151	4.23	162	4.54
Emergency	107	3.00	269	7.54
Inpatient	2033	57.01	2302	64.55
Laboratory	1	0.03	2303	64.58
Newborn	1	0.03	2304	64.61
Observation	21	0.59	2325	65.20
Observation / Short stay / 24 hr stay	77	2.16	2402	67.36
Obstetrics	4	0.11	2406	67.47
Other Specialty	59	1.65	2465	69.13
Outpatient	670	18.79	3135	87.91
Outpatient Surgery	211	5.92	3346	93.83
Preadmit	11	0.31	3357	94.14
Radiation Therapy	3	0.08	3360	94.22
Radiology	32	0.90	3392	95.12
Recurring	121	3.39	3513	98.51
Series	51	1.43	3564	99.94
Skilled Nursing Facility	1	0.03	3565	99.97
Urgent	1	0.03	3566	100.00

**Table.4.4** Variable analysis: Patient type

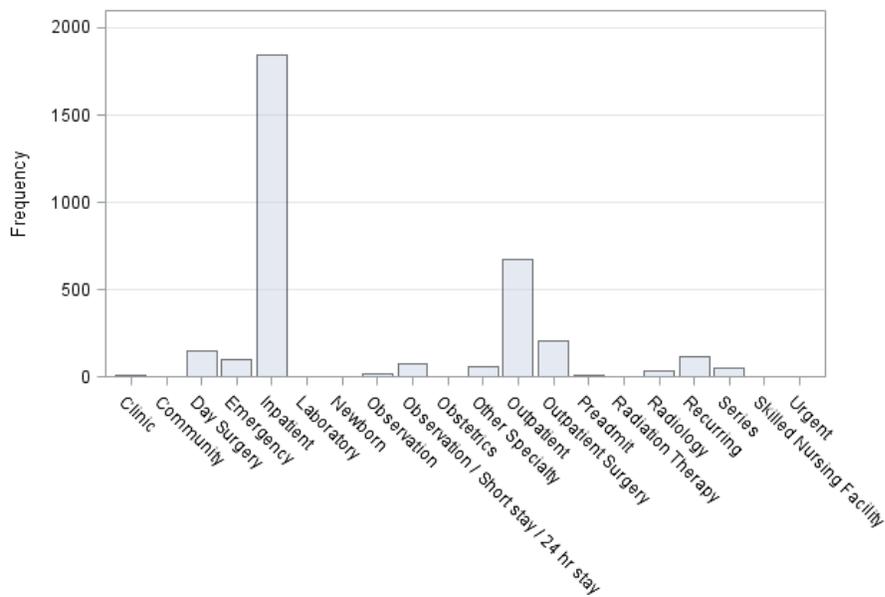


**Fig.4.12** Survival rate based on Patient type



**Fig.4.13** Bar Chart of Relative Percentages of Patient type

Figure 4.13 clearly points out “Inpatient” show higher likelihood of having ovarian cancer in comparison to other patient type. It also shows “Outpatient” are half of that of “Inpatient”. Figure 4.14 shows the distribution of survived patients with different patient type status.



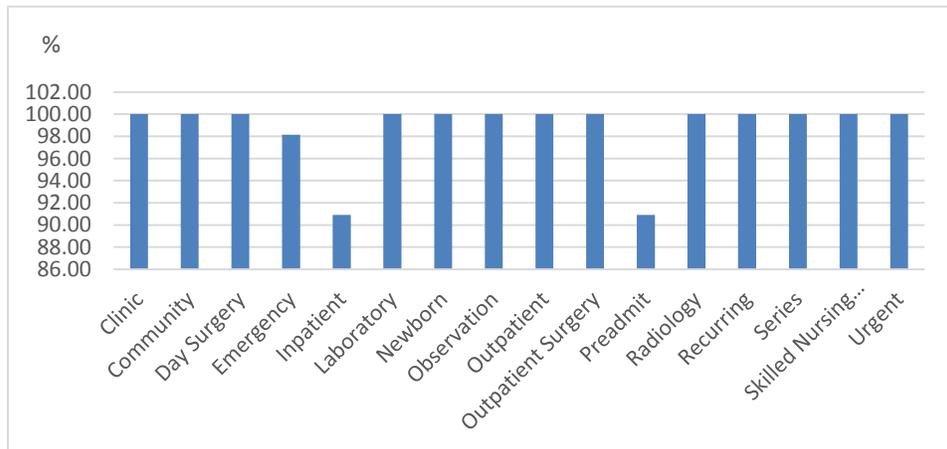
**Fig.4.14** Bar Chart of Patient type vs Survived =1

**Variable Analysis: Admission Source and Present On Admit**

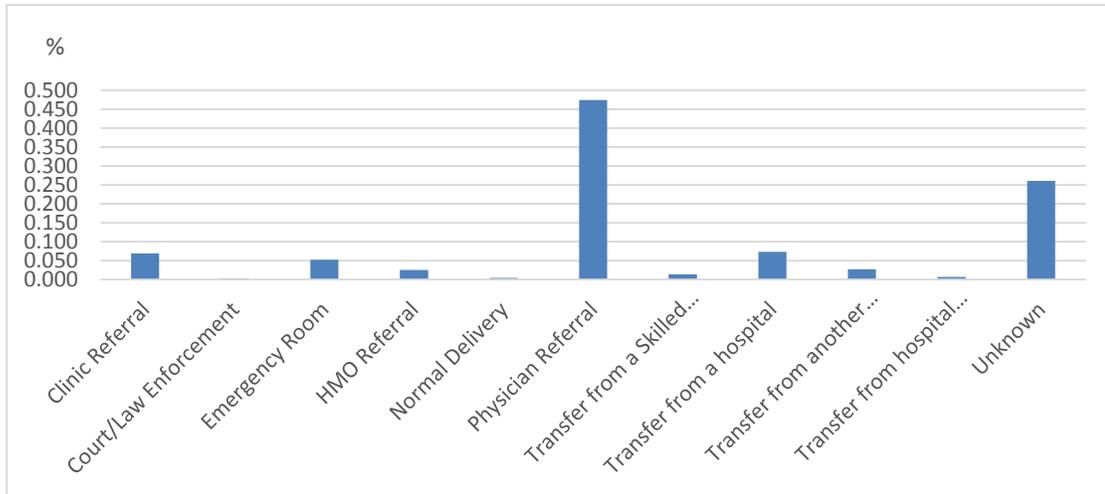
The admissions of these patients' visits are captured by the admission source used for the patients. It can be said that, 70 % of ovarian cancer patients are referred by physicians for further examination of their condition. On the other hand, there are 12% clinic referrals and 8 % emergency room visit, in the case of this study. The unknown in this case can be a case of missing data, or if the source of admission has not been captured while the data is collected. Also, 80% of the times the presence of ovarian cancer is unknown while the patient has been admitted to the healthcare organization. This is justified by the table 4.6:

admission_source_desc	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Clinic Referral	435	12.20	435	12.20
Court/Law Enforcement	2	0.06	437	12.25
Emergency Room	300	8.41	737	20.67
HMO Referral	8	0.22	745	20.89
Normal Delivery	1	0.03	746	20.92
Physician Referral	2496	69.99	3242	90.91
Transfer from a Skilled Nursing Facility (SNF)	8	0.22	3250	91.14
Transfer from a hospital	60	1.68	3310	92.82
Transfer from another health care facility	23	0.64	3333	93.47
Transfer from hospital inpt/same fac reslt in a sep claim	2	0.06	3335	93.52
Unknown	231	6.48	3566	100.00

**Table.4.5** Frequency distribution for different Admission sources

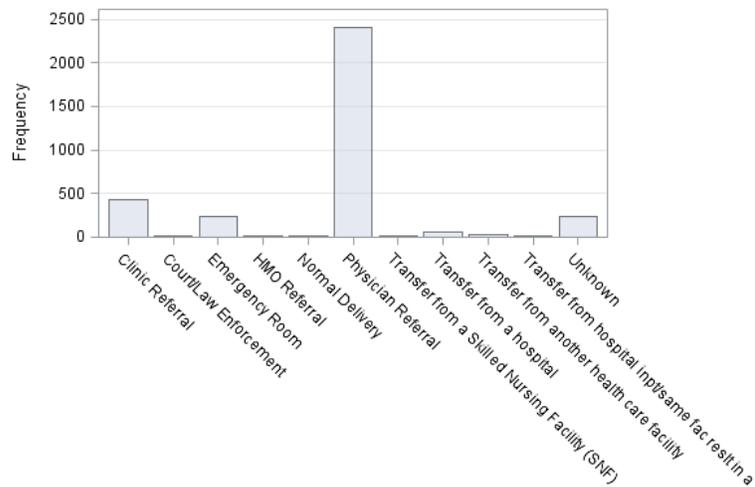


**Fig.4.15** Survival rate based on Admission sources



**Fig.4.16** Bar Chart of Relative Percentages of Admission sources

Figure 4.16 shows “Physician Referral” is the most frequent admission source of ovarian cancer, while “Clinic Referral” are just 11% of the total admission type. Figure 4.17 shows the distribution of survived patients with different admission sources.



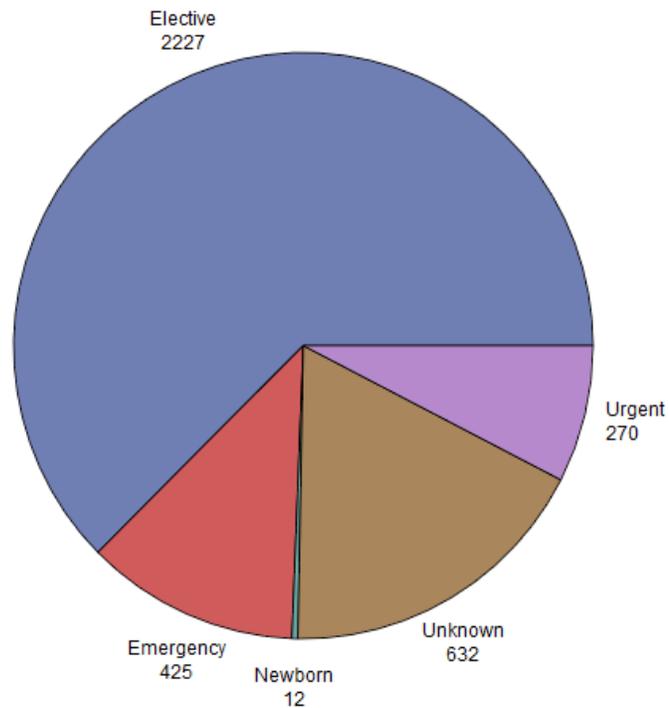
**Fig.4.17** Bar Chart of Admission Source vs Survived =1

PRESENT_ON_ADMIT_DESC	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Unknown	2862	80.26	2862	80.26
Yes	704	19.74	3566	100.00

**Table.4.6** Frequency distribution for Ovarian Cancer patients with Present on admit

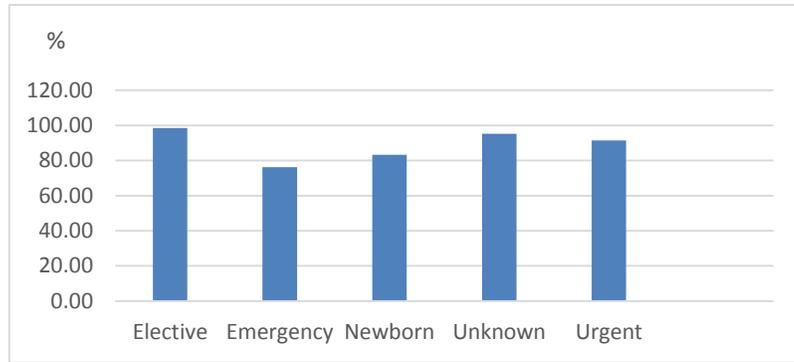
**Variable Analysis: Admission Type**

The pie chart (Fig. 4.18) shows distribution of the variable admission type that captures information on patient visits. It can be said that, 2,227 out of 3,566 of ovarian cancer patients are patients for whom the decision to admit has been made. On the other hand, there are 425 patients who are emergency case and 270 are urgent cases.

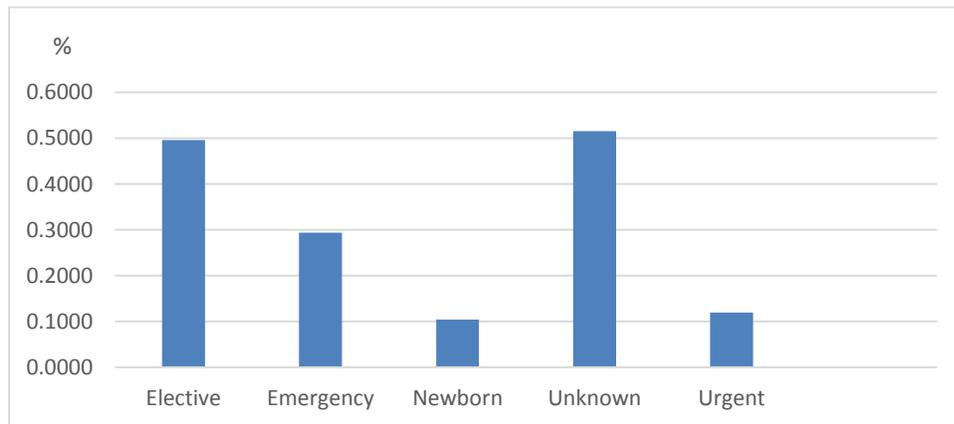


**Fig.4.18** Pie Chart of Admission type

Based on the survival rate patients admitted in “Emergency” show a survival rate of 70%, while patients admitted “Urgent” show 81% survival rate (Fig.4.19).

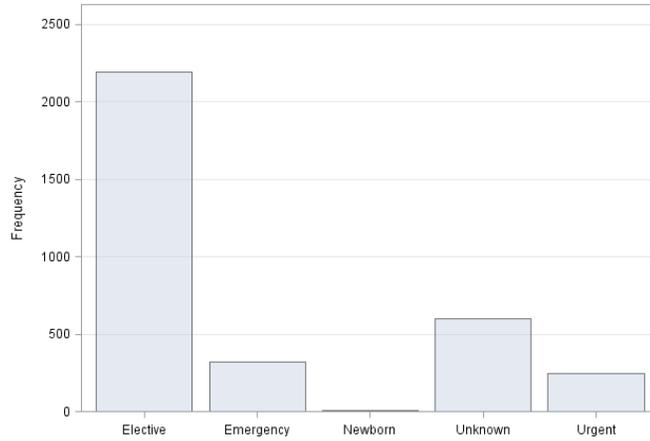


**Fig.4.19** Survival rate based on Admission type



**Fig.4.20** Bar Chart of Relative Percentages of Admission type

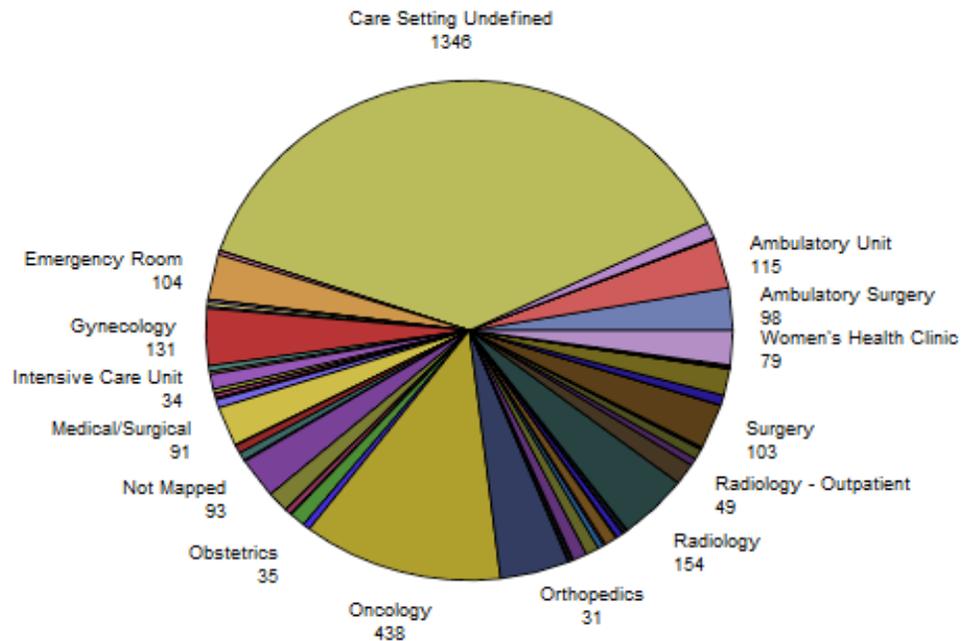
Figure 4.20 shows the decision to admit ovarian cancer patients has already made and this can also be said from previous graphs where “Physician Referral” is the most frequent admission. Physician apply for patient’s admission and thus the decision to admit such patients was made before they are admitted to the hospital. One of the reason why “Elective” admission type is higher as compared to other admission types. Figure 4.21 shows the distribution of survived patients with different admission type.



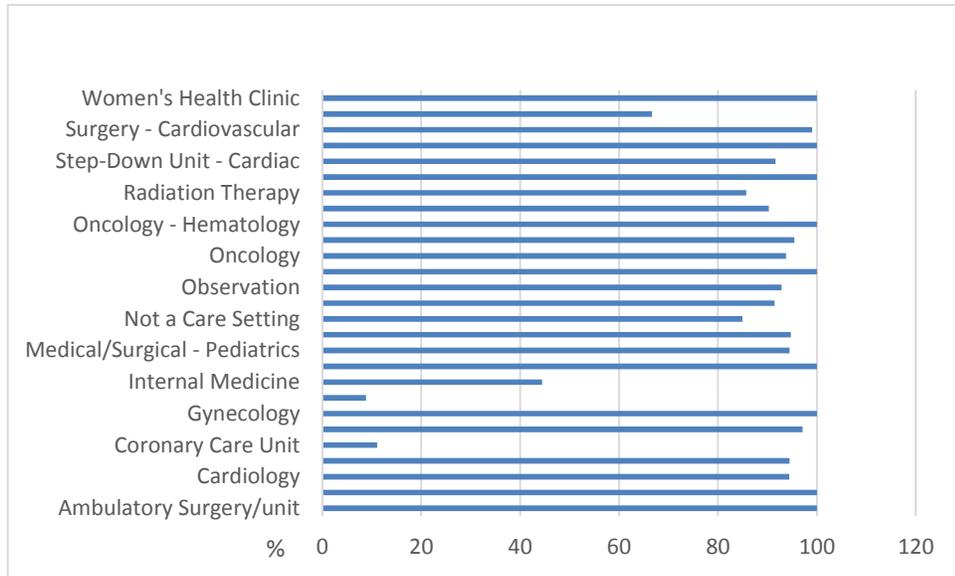
**Fig.4.21** Bar Chart of Admission type vs Survived =1

**Variable Analysis: Care Setting type**

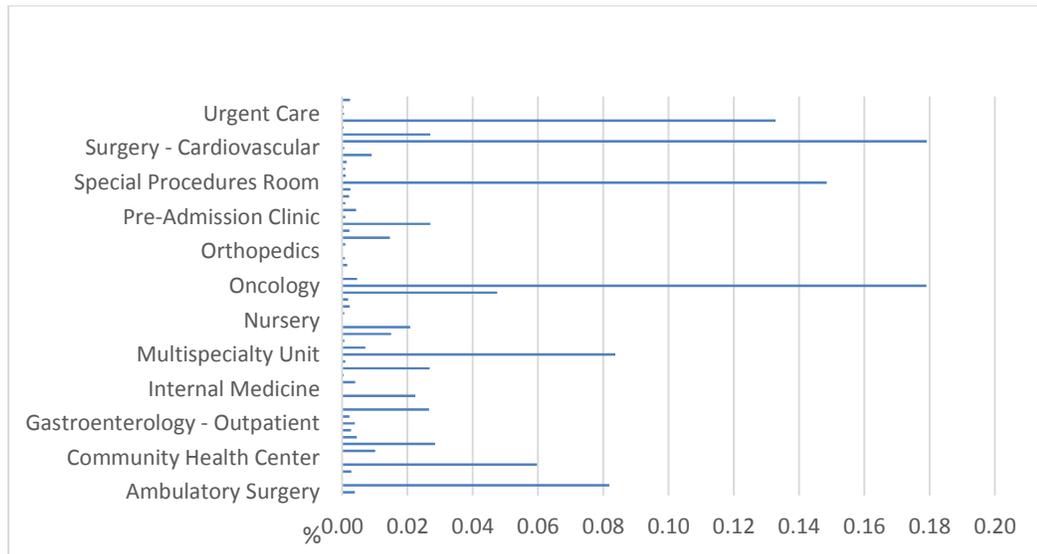
Pie chart Fig 4.22 shows various care-setting types. It can be said that, the care setting is undefined for 1,346 out of 3,566 patients, while 438 patients are sent to oncology. 104 patients are in the emergency room and 115 patients in ambulatory unit. Finally, 203 patients are sent for radiology.



**Fig.4.22** Pie Chart of Care-setting

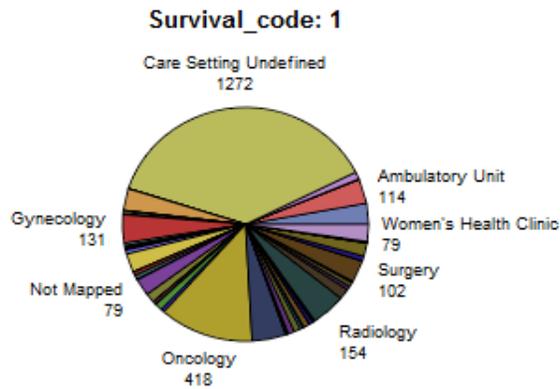


**Fig.4.23** Survival rate based on Care-setting



**Fig.4.24** Bar Chart of Relative Percentages of Care-setting

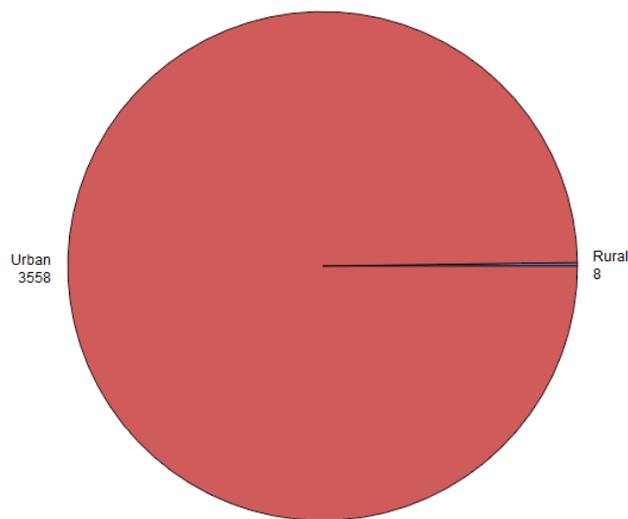
Fig 4.24 shows the maximum likelihood for care setting is for patients in “Surgery-Orthopedics” & “Oncology”. Patients in “Special Procedures Room” also have a high likelihood of having ovarian cancer. Figure 4.25 shows the distribution of survived patients with different care setting types.



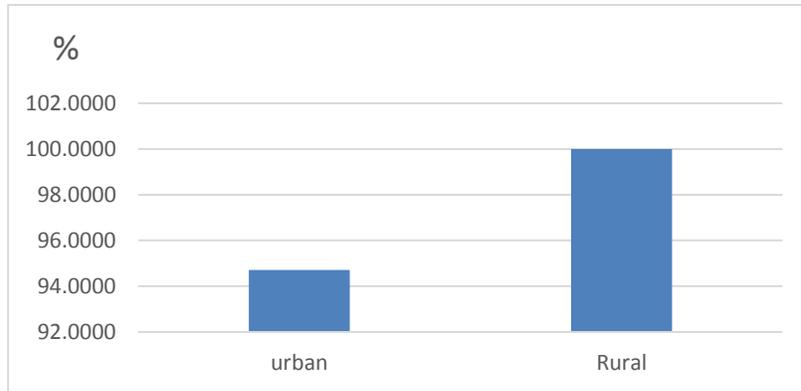
**Fig.4.25** Pie Chart of Relative Percentages of Care-setting vs Survived =1

#### Variable Analysis: Urban Rural Status

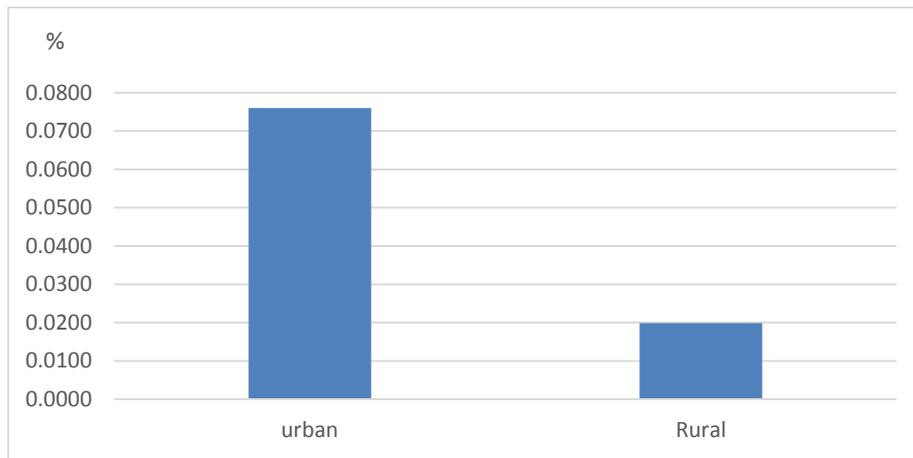
The dataset provides information of the Cerner’s facility being in the “Urban” region or the “Rural” region. It gives an indication of where the patient might belong to in terms of its “Urban” “Rural” status. Pie chart Fig.4.26 shows there are 3558 women out of 3566 women who belong to or visit the “Urban” facilities of Cerner hospital. While only 8 women belong to “Rural” facilities.



**Fig.4.26** Pie Chart of Urban Rural Status of Ovarian Cancer patients

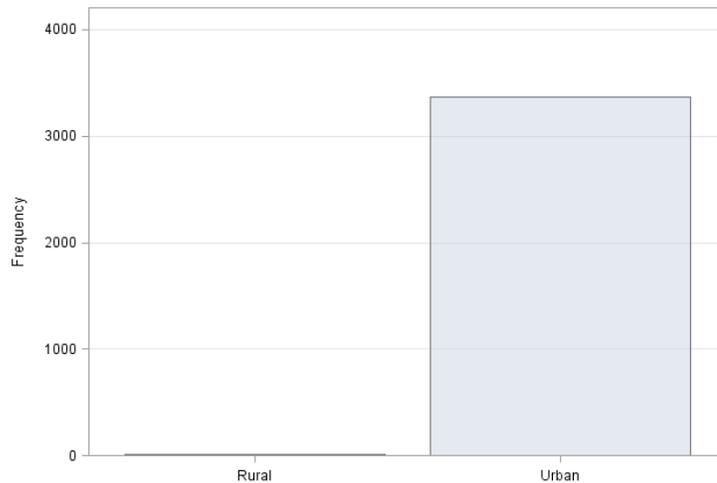


**Fig.4.27** Survival rate based on Urban vs Rural Status



**Fig.4.28** Bar Chart of Relative Percentages of Urban vs Rural Status

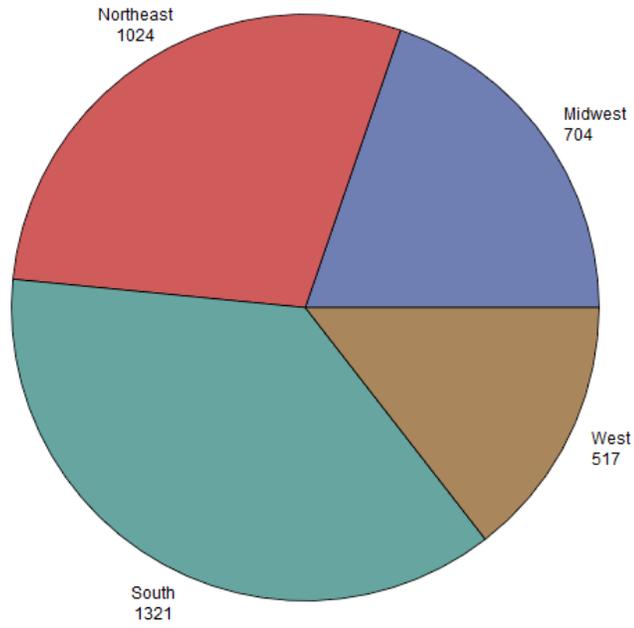
Figure 4.28 shows that number of patients admitted to urban areas is four times more than rural i.e. women in “Urban” area show higher likelihood of having ovarian cancer. Figure 4.29 shows the distribution of survived patients with different urban rural status. Based on survival rate women in “Rural” area show better results than “Urban” area.



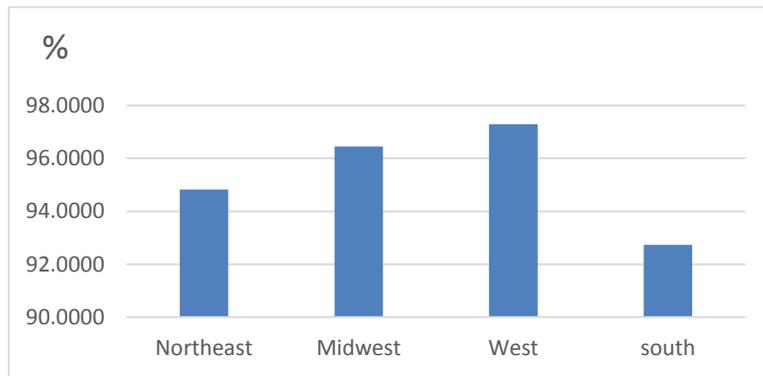
**Fig.4.29** Bar Chart of Urban Rural Status vs Survived =1

**Variable Analysis: Census Region**

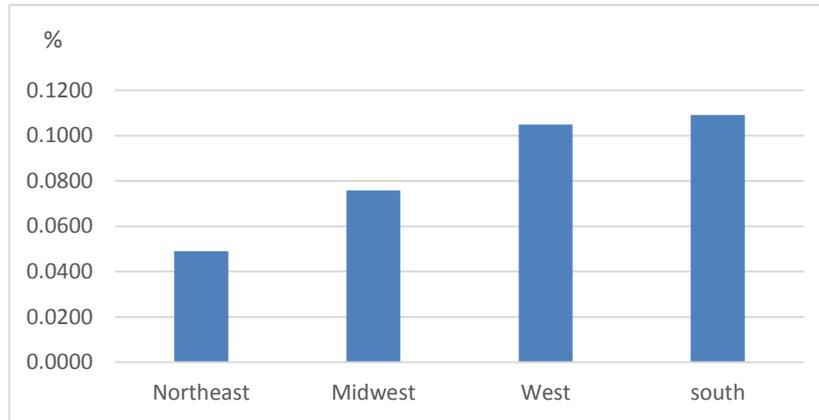
Pie chart Fig 4.30 deals with the distribution of ovarian cancer patients in the entire US where the Cerner facilities are located. Cerner dataset consists of four census regions in which the facility is located South (East South Central, South Atlantic, and West South Central), Midwest (East North Central, and West North Central), West (Mountain, and Pacific), and Northeast(New England, and Middle Atlantic). As the pie chart, shows there are 1321 patients out of 3566 that belong to the South (East South Central, South Atlantic, and West South Central) region. The next to this is the Midwest (East North Central, and West North Central) region with a population of 1024 women and the West (Mountain, and Pacific) region has the least number of ovarian cancer patients i.e. 517 women.



**Fig.4.30** Pie Chart of Census Region for Ovarian Cancer patients

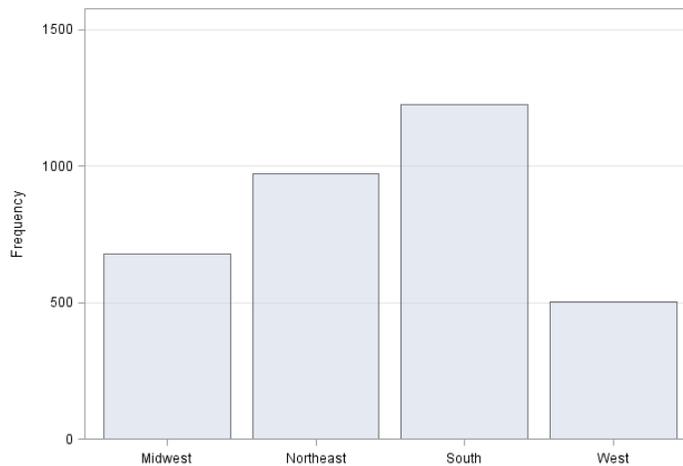


**Fig.4.31** Survival rate based on Census Region



**Fig.4.32** Bar Chart of Relative Percentages of Census Region

Figure 4.32 shows “South” region show higher likelihood of having ovarian cancer, which is more than double that of “Northeast” region. Figure 4.33 shows the distribution of survived patients with different census regions.



**Fig.4.33** Bar Chart of Census Region vs Survived =1

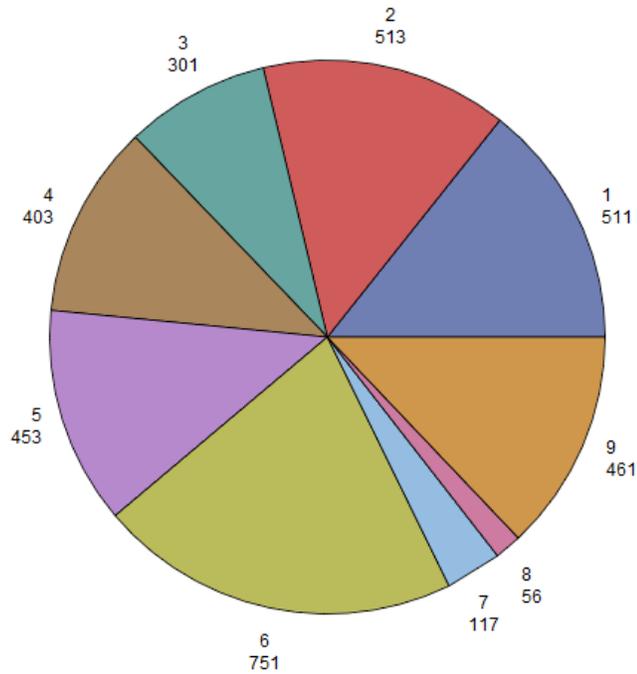
**Variable Analysis: Census Division**

For the Cerner’s dataset there are 9 census divisions in which the 4 regions are divided. South (East South Central, South Atlantic, and West South Central), Midwest (East North Central, and West North Central), West (Mountain, and Pacific), and Northeast (New England, and Middle

Atlantic). The census division is an indicator of which division each facility resides. The table 4.7 shows these divisions.

Census Division	States
1	New England (CT, ME, MA, NH, RI, VT)
2	Middle Atlantic (NJ, NY, PA)
3	West North Central (IA, KS, NE, MN, MO, ND, SD)
4	East North Central (IL, IN, MI, OH, WI)
5	East South (AL, KY, MS, TN)
6	South (DE, DC, FL, GA, MD, NC, SC, VA, WV)
7	West South Central (AR, LA, OK, TX)
8	Mountain (AZ, CO, ID, MT, NV, NM, UT, WY)
9	Pacific (AK, CA, HI, OR, WA)

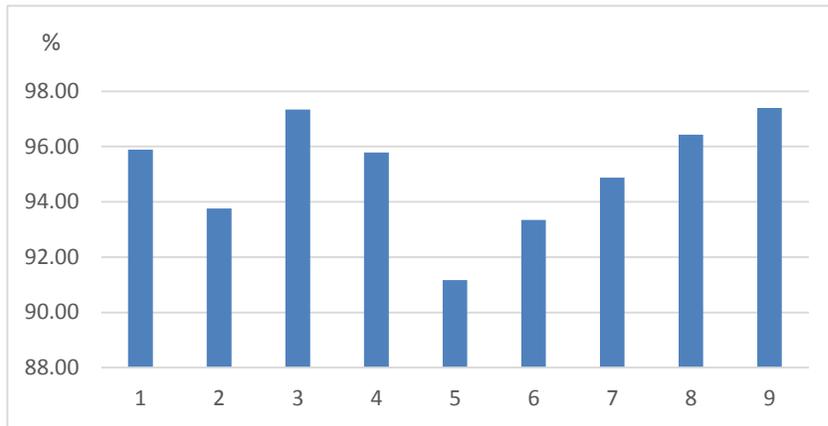
**Table.4.7** List of Census Divisions



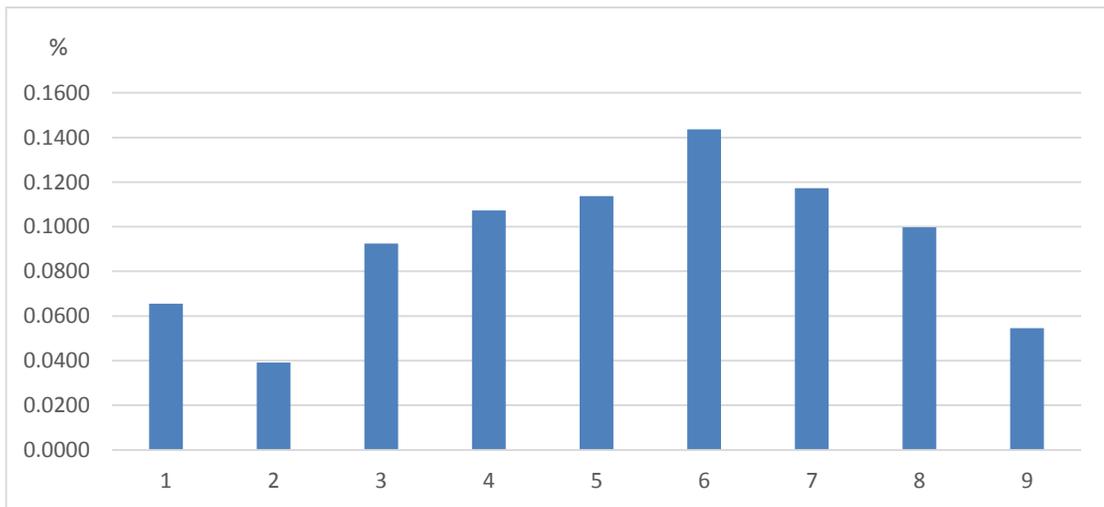
**Fig.4.34** Pie Chart of Census division for Ovarian Cancer patients

The pie chart fig.4.34 shows census division “6” South (DE, DC, FL, GA, MD, NC, SC, VA, WV) has the highest concentration of ovarian cancer patient i.e. 751. Census Division “2” has the second highest concentration of 513 patients in Middle Atlantic (NJ, NY, and PA). Census

Division “1” constitutes of 511 patients in New England (CT, ME, MA, NH, RI, VT) relatively 0.06%. While census division “9” constitutes of 461 patients from Pacific (AK, CA, HI, OR, WA). East South (AL, KY, MS, TN) census division of “5” constitutes 453 patients. “4” East North Central (IL, IN, MI, OH, WI) constitutes 403 patients and Mountain (AZ, CO, ID, MT, NV, NM, UT, and WY) census division “8” has the least count of ovarian cancer patients i.e. 56 out of 3566 patients.



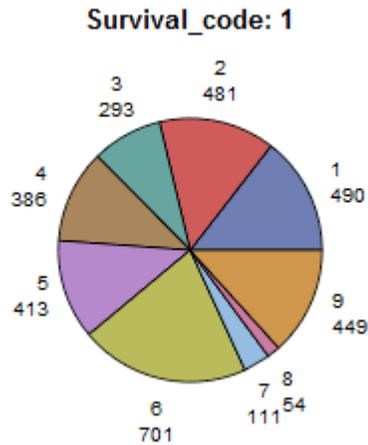
**Fig.4.35** Survival rate based on Census division



**Fig.4.36** Bar Chart of Relative Percentages of Census division

Figure 4.36 shows remarkably high number of patients in division 6( DE, DC, FL, GA, MD, NC, SC, VA, WV) i.e. Division 6 show higher likelihood of having ovarian cancer. The pattern in

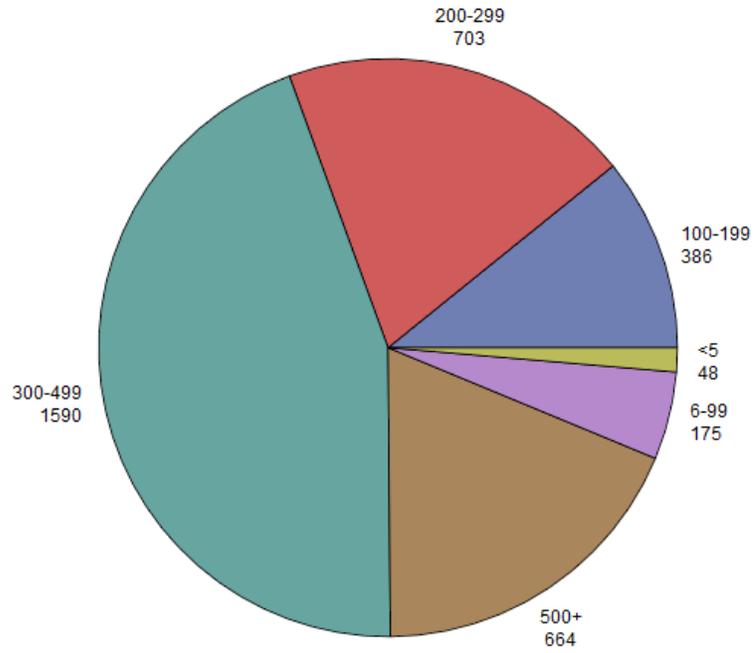
these states is reflective of their population. Figure 4.37 shows the distribution of survived patients with different census divisions.



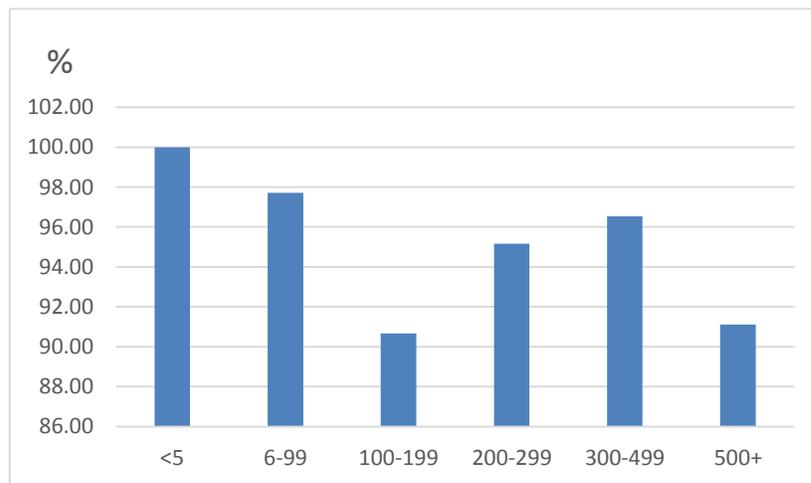
**Fig.4.37** Pie Chart of Census Division vs Survived =1

**Variable Analysis: Bed Size Range**

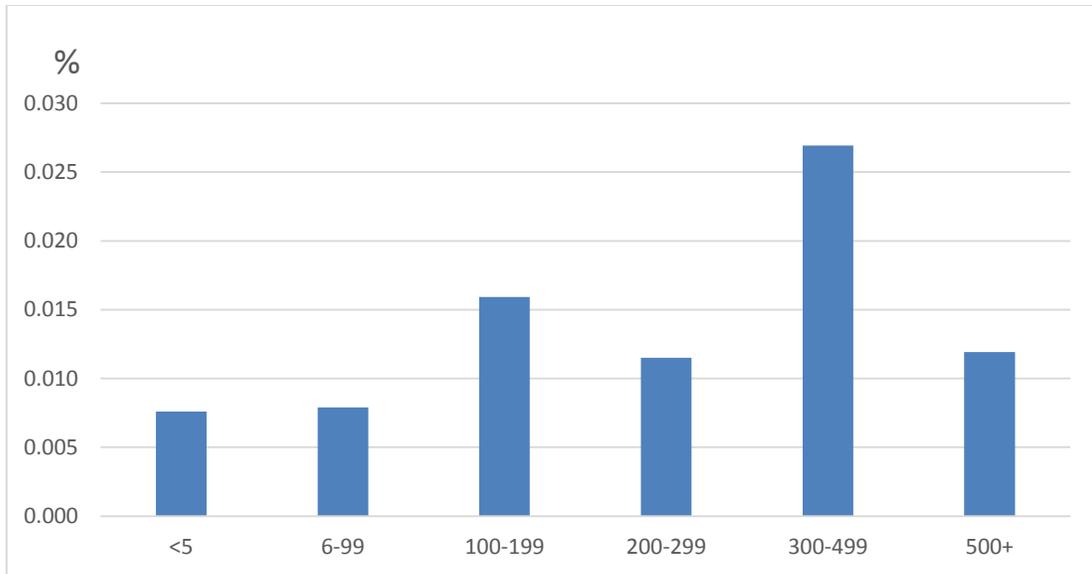
The dataset captures the information of the various categories of bed sizes that define the size of the hospital a patient visited. The pie chart fig.4.28 has the bed size categories of the hospital. The bed size categories vary from <5 to 500+ i.e. hospital consists of number of beds from <5 to 500+. The hospital having a bed size category of 300-499 has the most patient counts of 1590, while the second most frequent bed size category of the hospital is 200-299 with a patient count of 703 patients. 6-99 category has 175 patients. While the largest category 500+ has a patient count of 664 patients and the least size, < 5 has 48 patients.



**Fig.4.38** Pie Chart of Bed size range used for Ovarian Cancer patients

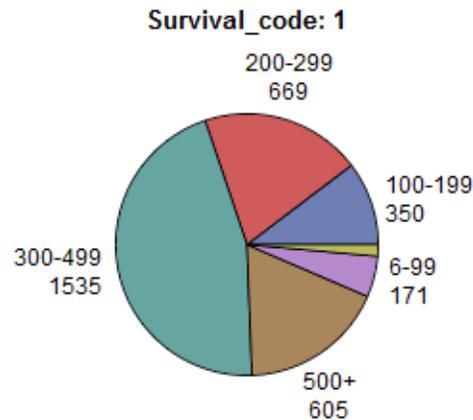


**Fig.4.39** Survival rate based on Bed size range



**Fig.4.40** Bar Chart of Relative Percentages of Bed size range

Figure 4.40 shows relative percentages of different bed size range. It is clear from the graph that patients visiting hospitals having bed size ‘300-499’ have maximum likelihood of having cancer followed by ‘100-199’. Figure 4.41 shows the distribution of survived patients with different bed size range.



**Fig.4.41** Pie Chart of Bed Size range vs Survived =1

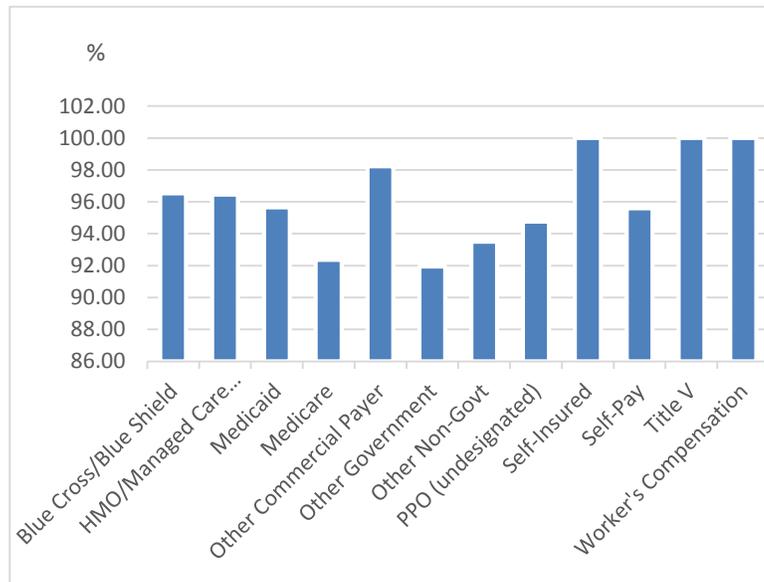
#### Variable Analysis: Payer Type

The payer type of each patient too have been captured in the dataset (Table 4.8). There are 36.17 % patient whose payer type is “Medicare”, followed by “Blue Cross/Blue Shield” with 14.55 %

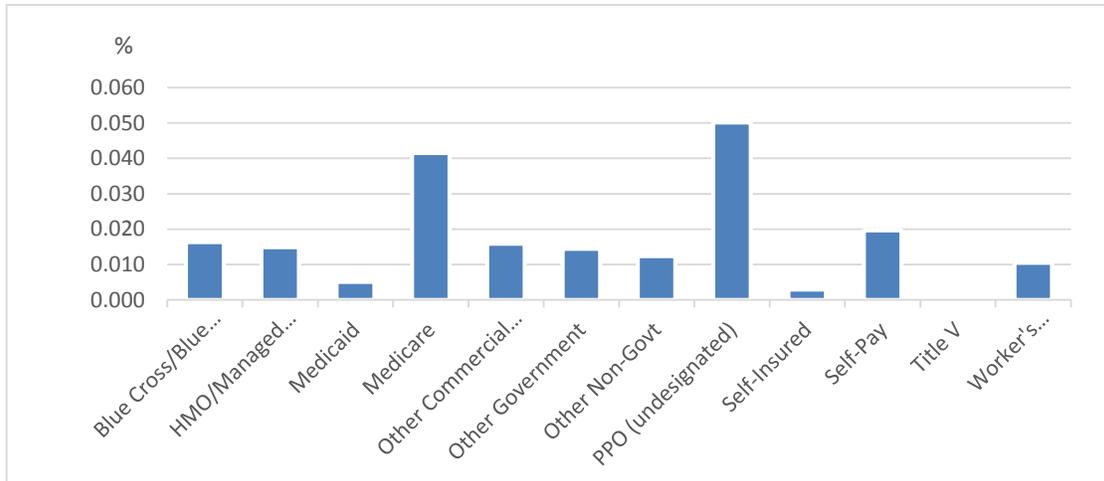
i.e. 519 patients. Only 3 patients or 0.08% patients are “Self-insured” and approximately, 0.17 % are “Worker’s Compensation” type of payers.

payer_code_desc	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Blue Cross/Blue Shield	519	14.55	519	14.55
CHAMPUS (Military dependents)	26	0.73	545	15.28
Free, Research	2	0.06	547	15.34
HMO/Managed Care (undesignated)	310	8.69	857	24.03
MIA	11	0.31	868	24.34
Medicaid	210	5.89	1078	30.23
Medicaid Managed Care (undesignated)	43	1.21	1121	31.44
Medicare	1290	36.17	2411	67.61
Medicare Managed Care (undesignated)	71	1.99	2482	69.60
Other Commercial Payer	450	12.62	2932	82.22
Other Government	87	2.44	3019	84.66
Other Non-Govt	123	3.45	3142	88.11
PPO (undesignated)	210	5.89	3352	94.00
Self-Insured	3	0.08	3355	94.08
Self-Pay	204	5.72	3559	99.80
Title V	1	0.03	3560	99.83
Worker’s Compensation	6	0.17	3566	100.00

**Table.4.8** Variable analysis: Payer type

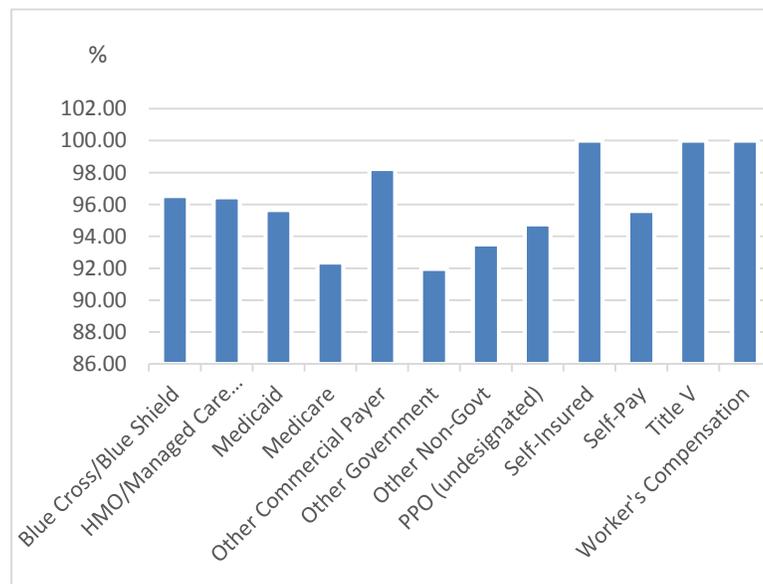


**Fig.4.42** Survival rate based on Payer type

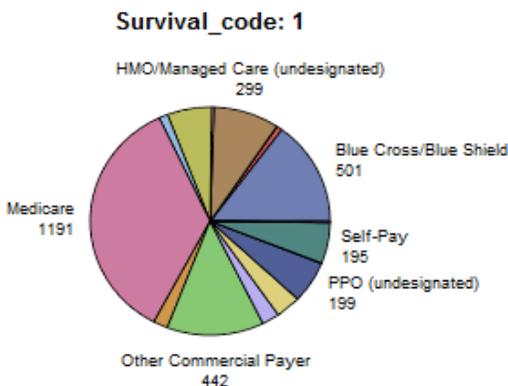


**Fig.4.42** Bar Chart of Relative Percentages of Payer type

Figure 4.42 shows relative percentages of different payer types. It is clear from the graph ‘PPO’ patients show higher likelihood of having ovarian cancer followed by ‘Medicare’. A very high number of patients also fall under the payer type “Self pay”. Figure 4.43 shows the distribution of survived patients with different payer type.



**Fig.4.43** Survival rate based on Payer type



**Fig.4.44** Pie Chart of Payer type vs Survived =1

**Variable Analysis: Total Charges**

Table 4.9 shows the total charges billed to the patients. As observed, the mean for total charges billed to ovarian cancer patients is \$15584.46. While it varies from \$0 to a \$738668.30, one of the reason for this variation could be the payer type of the patient.

Analysis Variable : total_charges									
Mean	Std Dev	Std Error	Variance	Minimum	Maximum	Mode	Range	N	N Miss
15584.46	31739.72	531.5111686	1007409700	0	738668.30	0	738668.30	3566	0

**Table.4.9** Variable analysis: Total charges billed

**Variable Analysis: Length of Stay**

Table 4.10 shows the length of stay of a patient. It clearly shows the average number of days a person is in hospital is 3.5 days.

Analysis Variable : Length_of_stay										
Mean	Std Dev	Std Error	Variance	Minimum	Maximum	Mode	Range	Sum	N	N Miss
3.5793606	13.2817039	0.2224145	176.4036580	0	602.0000000	0	602.0000000	12764.00	3566	0

**Table.4.10** Variable analysis: Length of Stay of a patient

## 4.2 Hypothesis Testing

### Hypothesis Testing: Survival vs Type of Healthcare Care Setting.

In the later part of this chapter, the findings from hypothesis testing are been discussed. All of the hypothesis have been tested at a significance level of 5 %. First hypothesis under consideration is, the null hypothesis that the survival of the patient is independent of healthcare type being an acute care setting. This null hypothesis was tested by looking at the Likelihood Ratio Chi-Square. As the evidence shows (Table 4.11), the Likelihood Ratio Chi-Square is greater than significance level of 5%, so the null hypothesis can't be rejected. Thus, it can be said that the survival of the patient is independent of healthcare type being an acute care setting. The strength of this association is a weak association, which can be stated by the Contingency Coefficient for this test.

$H_0$  = Survival of the patient is independent of the type of healthcare being an acute care setting.

$H_a$  = Survival of the patient is dependent of the type of healthcare being an acute care setting.

Statistic	DF	Value	Prob
Chi-Square	1	0.7822	0.3765
Likelihood Ratio Chi-Square	1	1.5196	0.2177
Continuity Adj. Chi-Square	1	0.0814	0.7754
Mantel-Haenszel Chi-Square	1	0.7820	0.3765
Phi Coefficient		0.0148	
Contingency Coefficient		0.0148	
Cramer's V		0.0148	

**Table.4.11** Statistics for table for Survival vs type of healthcare care setting.

From the cross tabulation table (Table 4.12) it is said that there are 99.61% women that were admitted to care setting type acute, while 0.39 % women are admitted to non-acute caresetting type. The percentage of women that survived who were admitted to care setting type acute are 99.59%.

Table of acute_care by Survival_code				
		Survival_code		Total
		Expired	Survived	
acute_care				
Acute	Frequency	188	3364	3552
	Percent	5.27	94.34	99.61
	Row Pct	5.29	94.71	
	Col Pct	100.00	99.59	
Non-Acute	Frequency	0	14	14
	Percent	0.00	0.39	0.39
	Row Pct	0.00	100.00	
	Col Pct	0.00	0.41	
Total	Frequency	188	3378	3566
	Percent	5.27	94.73	100.00

**Table.4.12** Cross Tabulation: Survival vs type of healthcare care setting

**Hypothesis Testing: Survival vs Bed Size Category of the Hospital**

The next null hypothesis is to find an association between the bed size category of the hospital and survival. The null hypothesis in this case is survival of the patient is independent of the bed size used for the patient. As evidence shows (Table 4.13), the Likelihood Ratio Chi-Square is less than the significance level of 5%, so the null hypothesis is rejected. Rejecting the null hypothesis means the alternative hypothesis is true. The alternative hypothesis in this case is survival of the patient is dependent on the bed size category of the hospital. Thus, there is an association between survival and the bed size category of the hospital. The variables show a weak association evidenced by Contingency Coefficient.

$H_0$  = Survival of the patient is independent of the bed size category of the hospital.

$H_a$  = Survival of the patient is dependent of the bed size category of the hospital.

From the cross tabulation table (Table 4.13) it is said that there are 44.59% women that were admitted to hospital with bed size range of 300-499, while 19.71 % women are admitted to

hospital with bed size range of 200-299. The percentage of women that survived who were admitted to hospital with bed size range of 300-499 are 45.44%, while 29.26% were expired. 19.80 % women survived out of all the patients that were admitted to hospital with bed size range of 200-299, while 18.09% women expired in that category. <5 category constitutes 1.35% women and 100-199, 10.82 % women.

Table of bed_size_range by Survival_code				
		Survival_code		Total
		Expired	Survived	
bed_size_range				
100-199	Frequency	36	350	386
	Percent	1.01	9.81	10.82
	Row Pct	9.33	90.67	
	Col Pct	19.15	10.36	
200-299	Frequency	34	669	703
	Percent	0.95	18.76	19.71
	Row Pct	4.84	95.16	
	Col Pct	18.09	19.80	
300-499	Frequency	55	1535	1590
	Percent	1.54	43.05	44.59
	Row Pct	3.46	96.54	
	Col Pct	29.26	45.44	
500+	Frequency	59	605	664
	Percent	1.65	16.97	18.62
	Row Pct	8.89	91.11	
	Col Pct	31.38	17.91	
6-99	Frequency	4	171	175
	Percent	0.11	4.80	4.91
	Row Pct	2.29	97.71	
	Col Pct	2.13	5.06	
<5	Frequency	0	48	48
	Percent	0.00	1.35	1.35
	Row Pct	0.00	100.00	
	Col Pct	0.00	1.42	
Total	Frequency	188	3378	3566
	Percent	5.27	94.73	100.00

**Table.4.13** Cross Tabulation: Bed size range vs Survival code

Statistic	DF	Value	Prob
Chi-Square	5	46.5937	<.0001
Likelihood Ratio Chi-Square	5	46.2356	<.0001
Mantel-Haenszel Chi-Square	1	2.4768	0.1155
Phi Coefficient		0.1143	
Contingency Coefficient		0.1136	
Cramer's V		0.1143	

**Table.4.14** Statistics for table of Bed size range by Survival code

#### **Hypothesis Testing: Survival vs Census Region**

The next test looks at whether survival and the census region of where the patient belongs have any association. The null hypothesis is that survival of the patient is independent of the census region of where the patient lives. The Likelihood Ratio Chi-Square shows (Table 4.16) the chi-square probability to be less than significance level of 5%, thus the null hypothesis was rejected. This implies the alternative hypothesis is true; survival of the patient is dependent on the census region where the patient lives. It can be thus, concluded that there is an association between survival and census region. The strength of this association is a weak association, this can be said based on the Contingency Coefficient for this test.

$H_0$  = Survival of the patient is independent of the census region of where the patient lives.

$H_a$  = Survival of the patient is dependent on the census region of where the patient lives.

Cross tabulation table shows (Table 4.15) it is said that there are 37.04% women that belong to South region, while 28.72 % women are from Northeast. The percentage of women that survived who are from the South region are 36.26%, while 51.06% were expired. 28.74% women survived were from Northeast, while 28.19% women expired in that category. 19.74% women are in Midwest 20.10 % women survived and 13.30% women expired. While, there are 14.50% women from West out of which 14.89% survived and 7.45 % expired.

Table of census_region by Survival_code				
		Survival_code		Total
		Expired	Survived	
census_region				
Midwest	Frequency	25	679	704
	Percent	0.70	19.04	19.74
	Row Pct	3.55	96.45	
	Col Pct	13.30	20.10	
Northeast	Frequency	53	971	1024
	Percent	1.49	27.23	28.72
	Row Pct	5.18	94.82	
	Col Pct	28.19	28.74	
South	Frequency	96	1225	1321
	Percent	2.69	34.35	37.04
	Row Pct	7.27	92.73	
	Col Pct	51.06	36.26	
West	Frequency	14	503	517
	Percent	0.39	14.11	14.50
	Row Pct	2.71	97.29	
	Col Pct	7.45	14.89	
Total	Frequency	188	3378	3566
	Percent	5.27	94.73	100.00

**Table.4.15** Cross Tabulation: Census Region vs Survival code

Statistic	DF	Value	Prob
Chi-Square	3	21.5296	<.0001
Likelihood Ratio Chi-Square	3	22.3948	<.0001
Mantel-Haenszel Chi-Square	1	0.8601	0.3537
Phi Coefficient		0.0777	
Contingency Coefficient		0.0775	
Cramer's V		0.0777	

**Table.4.16** Statistics for table of Census Region by Survival code

**Hypothesis Testing: Survival vs Race**

Another test looks at the association between survival and the race of the patient. The null hypothesis of survival of the patient is independent of the race to which the patient belongs.

Likelihood Ratio Chi-Square show (Table 4.17) the chi-square is less than the significance level of 5%, so the null hypothesis is rejected. Rejecting the null hypothesis means the alternative hypothesis is true. The alternative hypothesis in this case is survival of the patient is dependent on the race to which the patient belongs. Thus, it can be inferred that there is a relationship between the between survival and the race of the patient. Yet, the variables show a weak association as shown by Contingency Coefficient.

$H_0$  = Survival of the patient is independent of the race to which the patient belongs

$H_a$  = Survival of the patient is dependent on the race to which the patient belongs

Statistic	DF	Value	Prob
Chi-Square	7	15.0855	0.0349
Likelihood Ratio Chi-Square	7	19.4800	0.0068
Mantel-Haenszel Chi-Square	1	2.0561	0.1516
Phi Coefficient		0.0650	
Contingency Coefficient		0.0649	
Cramer's V		0.0650	

**Table.4.17** Statistics for table of Race by Survival code

Table of race by Survival_code				
		Survival_code		Total
		Expired	Survived	
race				
African American	Frequency	31	395	426
	Percent	0.87	11.08	11.95
	Row Pct	7.28	92.72	
	Col Pct	16.49	11.69	
Asian	Frequency	2	72	74
	Percent	0.06	2.02	2.08
	Row Pct	2.70	97.30	
	Col Pct	1.06	2.13	
Caucasian	Frequency	148	2687	2835
	Percent	4.15	75.35	79.50
	Row Pct	5.22	94.78	
	Col Pct	78.72	79.54	
Hispanic	Frequency	0	31	31
	Percent	0.00	0.87	0.87
	Row Pct	0.00	100.00	
	Col Pct	0.00	0.92	
Native American	Frequency	0	7	7
	Percent	0.00	0.20	0.20
	Row Pct	0.00	100.00	
	Col Pct	0.00	0.21	
Other	Frequency	1	126	127
	Percent	0.03	3.53	3.56
	Row Pct	0.79	99.21	
	Col Pct	0.53	3.73	
Pacific Islander	Frequency	0	8	8
	Percent	0.00	0.22	0.22
	Row Pct	0.00	100.00	
	Col Pct	0.00	0.24	
Unknown	Frequency	6	52	58
	Percent	0.17	1.46	1.63
	Row Pct	10.34	89.66	
	Col Pct	3.19	1.54	
Total	Frequency	188	3378	3566
	Percent	5.27	94.73	100.00

**Table.4.18** Cross Tabulation: Race vs Survival code

Cross tabulation table shows (Table 4.18) it is said that there are 79.50% women that are Caucasian, while 11.95 % women are African American. The percentage of women that survived who are Caucasian 79.54%, while 78.72% were expired. 11.69 % women survived were African American, while 16.49% women expired in that category. Hispanic women constituted 0.87 %

out of these 0.92% women survived. While, there are 2.08% women are Asian out of which 2.13% survived and 1.06 % expired.

### Hypothesis Testing: Survival vs Marital Status

The next null hypothesis is to find association between the marital status of the patient and survival. The null hypothesis in this case is the survival of the patient is independent of marital status of the patient. As evidence shows (Table 4.19), the Likelihood Ratio Chi-Square is less than the significance level of 5%, so the null hypothesis is rejected. This implies the alternative hypothesis is true, survival of the patient is dependent on marital status of the patient. It can be thus, concluded that there is an association between survival and marital status.

$H_0$  = Survival of the patient is independent of marital status of the patient.

$H_a$  = Survival of the patient is dependent on marital status of the patient.

Statistic	DF	Value	Prob
Chi-Square	6	27.0018	0.0001
Likelihood Ratio Chi-Square	6	24.8297	0.0004
Mantel-Haenszel Chi-Square	1	3.6710	0.0554
Phi Coefficient		0.0870	
Contingency Coefficient		0.0867	
Cramer's V		0.0870	

**Table.4.19** Statistics for table of Marital Status by Survival code

Cross tabulation table shows (Table 4.20) it is said that there are 46.21% women that are Married, while 16.63 % women are Single. The percentage of women that survived who are Married 46.60%, while 39.36% were expired. 17.02 % women survived were single, while 9.57% women expired in that category. Widowed women constituted 14.27 % out of these 13.71% widowed women survived and 24.47% widowed women expired. While, there are 0.08% are women who have life partner and 1.15% are legally separated women.

Table of marital_status by Survival_code				
		Survival_code		Total
		Expired	Survived	
marital_status				
Divorced	Frequency	18	297	315
	Percent	0.50	8.33	8.83
	Row Pct	5.71	94.29	
	Col Pct	9.57	8.79	
Legally Separated	Frequency	5	36	41
	Percent	0.14	1.01	1.15
	Row Pct	12.20	87.80	
	Col Pct	2.66	1.07	
Life Partner	Frequency	0	3	3
	Percent	0.00	0.08	0.08
	Row Pct	0.00	100.00	
	Col Pct	0.00	0.09	
Married	Frequency	74	1574	1648
	Percent	2.08	44.14	46.21
	Row Pct	4.49	95.51	
	Col Pct	39.36	46.60	
Single	Frequency	18	575	593
	Percent	0.50	16.12	16.63
	Row Pct	3.04	96.96	
	Col Pct	9.57	17.02	
Unknown	Frequency	27	430	457
	Percent	0.76	12.06	12.82
	Row Pct	5.91	94.09	
	Col Pct	14.36	12.73	
Widowed	Frequency	46	463	509
	Percent	1.29	12.98	14.27
	Row Pct	9.04	90.96	
	Col Pct	24.47	13.71	
Total	Frequency	188	3378	3566
	Percent	5.27	94.73	100.00

**Table.4.20** Cross Tabulation: Marital Status vs Survival code

**Hypothesis Testing: Survival vs Payer Type**

Much research has been done to find patterns or associations between the insurance a patient buys and its impact on the survival of the patient. To test the issue the null hypothesis is set to the survival of the patient, which is independent of payer type of the patient.

$H_0$  = Survival of the patient is independent of payer type of the patient.

$H_a$  = Survival of the patient is dependent on payer type of the patient.

As evidence shows (Table 4.21), the chi-square is less than the significance level of 5%, so the null hypothesis was rejected. Rejecting the null hypothesis means the alternative hypothesis is true. The alternative hypothesis in this case is survival of the patient is dependent on payer type of the patient. Thus, there is an association between survival and the payer type of the patient.

Statistic	DF	Value	Prob
Chi-Square	16	36.1337	0.0028
Likelihood Ratio Chi-Square	16	39.7461	0.0008
Mantel-Haenszel Chi-Square	1	1.5178	0.2180
Phi Coefficient		0.1007	
Contingency Coefficient		0.1002	
Cramer's V		0.1007	

**Table.4.21** Statistics for table of Payer type by Survival code

From the cross tabulation table (Table 4.22) it is said that there are 7.12% women that belong have Medicaid as their payer, while 38.16 % women have Medicaid as their payer type. The percentage of women that survived who had Medicaid payer were 7.16%, while 5.86% were expired. 37.21 % women survived who had Medicare payer, while 55.32 % women expired in that category. Blue cross/ Blue Shield constituted 14.55 % of the total payer types out of these 14.83 % women survived and 9.57% expired. Self-pay has 5.72 % of total out of these 5.77% survived and 4.79% expired.

Table of payer_code_desc by Survival_code				
payer_code_desc		Survival_code		Total
		Expired	Survived	
Blue Cross/Blue Shield	Frequency	18	501	519
	Percent	0.50	14.05	14.55
	Row Pct	3.47	96.53	
	Col Pct	9.57	14.83	
CHAMPUS (Military dependents)	Frequency	1	25	26
	Percent	0.03	0.70	0.73
	Row Pct	3.85	96.15	
	Col Pct	0.53	0.74	
Free, Research	Frequency	0	2	2
	Percent	0.00	0.06	0.06
	Row Pct	0.00	100.00	
	Col Pct	0.00	0.06	
HMO/Managed Care (undesignated)	Frequency	11	299	310
	Percent	0.31	8.38	8.69
	Row Pct	3.55	96.45	
	Col Pct	5.85	8.85	
MIA	Frequency	0	11	11
	Percent	0.00	0.31	0.31
	Row Pct	0.00	100.00	
	Col Pct	0.00	0.33	
Medicaid	Frequency	8	202	210
	Percent	0.22	5.66	5.89
	Row Pct	3.81	96.19	
	Col Pct	4.26	5.98	
Medicaid Managed Care (undesignated)	Frequency	3	40	43
	Percent	0.08	1.12	1.21
	Row Pct	6.98	93.02	
	Col Pct	1.60	1.18	
Medicare	Frequency	99	1191	1290
Medicare	Percent	2.78	33.40	36.17
	Row Pct	7.67	92.33	
	Col Pct	52.66	35.26	

Medicare Managed Care (undesignated)	Frequency	5	66	71
	Percent	0.14	1.85	1.99
	Row Pct	7.04	92.96	
	Col Pct	2.66	1.95	
Other Commercial Payer	Frequency	8	442	450
	Percent	0.22	12.39	12.62
	Row Pct	1.78	98.22	
	Col Pct	4.26	13.08	
Other Government	Frequency	7	80	87
	Percent	0.20	2.24	2.44
	Row Pct	8.05	91.95	
	Col Pct	3.72	2.37	
Other Non-Govt	Frequency	8	115	123
	Percent	0.22	3.22	3.45
	Row Pct	6.50	93.50	
	Col Pct	4.26	3.40	
PPO (undesignated)	Frequency	11	199	210
	Percent	0.31	5.58	5.89
	Row Pct	5.24	94.76	
	Col Pct	5.85	5.89	
Self-Insured	Frequency	0	3	3
	Percent	0.00	0.08	0.08
	Row Pct	0.00	100.00	
	Col Pct	0.00	0.09	
Self-Pay	Frequency	9	195	204
	Percent	0.25	5.47	5.72
	Row Pct	4.41	95.59	
	Col Pct	4.79	5.77	
Title V	Frequency	0	1	1
	Percent	0.00	0.03	0.03
	Row Pct	0.00	100.00	
	Col Pct	0.00	0.03	
Worker's Compensation	Frequency	0	6	6
	Percent	0.00	0.17	0.17
	Row Pct	0.00	100.00	
	Col Pct	0.00	0.18	
Total	Frequency	188	3378	3566
	Percent	5.27	94.73	100.00

**Table.4.22** Cross Tabulation: Payer type vs Survival code

#### **Hypothesis Testing: Survival vs Age**

Another hypothesis under consideration is the null hypothesis that the survival of the patient is independent of the age of the patient. This null hypothesis was tested by looking at the Likelihood Ratio Chi-Square. As evidence shows (Table 4.23), the Likelihood Ratio Chi-Square is less than the significance level of 5%, so the null hypothesis is rejected. This implies the alternative

hypothesis is true, survival of the patient is dependent on age of the patient. It can be thus, concluded that there is an association between survival and age.

$H_0$  = Survival of the patient is independent of the type of the age of the patient.

$H_a$  = Survival of the patient is dependent of the type of the age of the patient.

Statistic	DF	Value	Prob
Chi-Square	8	43.4604	<.0001
Likelihood Ratio Chi-Square	8	41.2084	<.0001
Mantel-Haenszel Chi-Square	1	33.8690	<.0001
Phi Coefficient		0.1104	
Contingency Coefficient		0.1097	
Cramer's V		0.1104	

**Table.4.23** Statistics for table for Survival vs age of patient

Table of age_in_years by Survival_code				
		Survival_code		Total
		Expired	Survived	
age_in_years				
Less than or equal to 17	Frequency	0	37	37
	Percent	0.00	1.04	1.04
	Row Pct	0.00	100.00	
	Col Pct	0.00	1.10	
16 to 27	Frequency	1	67	68
	Percent	0.03	1.88	1.91
	Row Pct	1.47	98.53	
	Col Pct	0.53	1.98	
28 to 37	Frequency	4	142	146
	Percent	0.11	3.98	4.09
	Row Pct	2.74	97.26	
	Col Pct	2.13	4.20	
38 to 47	Frequency	9	378	387
	Percent	0.25	10.60	10.85
	Row Pct	2.33	97.67	
	Col Pct	4.79	11.19	
48 to 57	Frequency	35	752	787
	Percent	0.98	21.09	22.07
	Row Pct	4.45	95.55	
	Col Pct	18.62	22.26	
58 to 67	Frequency	48	930	978
	Percent	1.35	26.08	27.43
	Row Pct	4.91	95.09	
	Col Pct	25.53	27.53	
68 to 77	Frequency	48	736	784
	Percent	1.35	20.64	21.99
	Row Pct	6.12	93.88	
	Col Pct	25.53	21.79	
78 to 87	Frequency	37	297	334
	Percent	1.04	8.33	9.37
	Row Pct	11.08	88.92	
	Col Pct	19.68	8.79	
88 to 97	Frequency	6	39	45
	Percent	0.17	1.09	1.26
	Row Pct	13.33	86.67	
	Col Pct	3.19	1.15	
Total	Frequency	188	3378	3566
	Percent	5.27	94.73	100.00

**Table.4.24** Cross Tabulation: Survival vs age of patient

From the cross tabulation table (Table 4.24) it is said that there are 27.43% women that belong to the age group 58 to 67 years. While the age group 48 to 57 constitutes 22.07% of the total patients and 1.26 % belong to 88 to 97. The age group 58 to 67 years has 27.53% patients survived, while 25.53% that expired. Age group of 68 to 77 has 25.53% women who expired, while 21.79 % women who survived.

### Hypothesis Testing: Survival vs Length of Stay

The next null hypothesis is to find an association between the length of stay in the facility by the patient and survival. The null hypothesis in this case is survival of the patient is independent of the length of stay in the facility by the patient. This null hypothesis was tested by looking at the Likelihood Ratio Chi-Square. As evidence shows (Table 4.25), the Likelihood Ratio Chi-Square is less than the significance level of 5%, so the null hypothesis is rejected. This implies the alternative hypothesis is true, survival of the patient is dependent on length of stay of the patient. It can be thus, concluded that there is an association between survival and length of stay of the patient.

$H_0$  = Survival of the patient is independent of the length of stay in the facility by the patient.

$H_a$  = Survival of the patient is dependent of the length of stay in the facility by the patient.

Statistic	DF	Value	Prob
Chi-Square	5	301.7572	<.0001
Likelihood Ratio Chi-Square	5	141.7485	<.0001
Mantel-Haenszel Chi-Square	1	236.7944	<.0001
Phi Coefficient		0.2909	
Contingency Coefficient		0.2793	
Cramer's V		0.2909	

**Table.4.25** Statistics for table for Survival vs length of stay in the facility by the patient.

Table of Length_of_stay by Survival_code					
		Survival_code		Total	
		Expired	Survived		
Length_of_stay					
	Less than or equal to 10	Frequency	130	3221	3351
		Percent	3.65	90.33	93.97
		Row Pct	3.88	96.12	
		Col Pct	69.15	95.35	
11 to 19		Frequency	33	130	163
		Percent	0.93	3.65	4.57
		Row Pct	20.25	79.75	
		Col Pct	17.55	3.85	
21 to 30		Frequency	14	17	31
		Percent	0.39	0.48	0.87
		Row Pct	45.16	54.84	
		Col Pct	7.45	0.50	
31 to 40		Frequency	3	5	8
		Percent	0.08	0.14	0.22
		Row Pct	37.50	62.50	
		Col Pct	1.60	0.15	
41 to 50		Frequency	6	1	7
		Percent	0.17	0.03	0.20
		Row Pct	85.71	14.29	
		Col Pct	3.19	0.03	
50 or more		Frequency	2	4	6
		Percent	0.06	0.11	0.17
		Row Pct	33.33	66.67	
		Col Pct	1.06	0.12	
Total		Frequency	188	3378	3566
		Percent	5.27	94.73	100.00

**Table.4.26** Cross Tabulation: Survival vs length of stay

From the cross tabulation table (Table 4.26) it can be said that out of patients who stay in facility for 11 to 19 days 20.25% patients expired, while 79.75% patients survived. When patients stay for 16 to 27 days in the facility 98.70% women survived on the other hand only 1.30% women expired.

#### **Hypothesis Testing: Survival vs Total Charges**

Another hypothesis under consideration is the null hypothesis that the survival of the patient is independent of the total charges a patient is charged. This null hypothesis was tested by looking at the Likelihood Ratio Chi-Square. As the evidence shows (Table 4.27), the Likelihood Ratio Chi-Square is less than significance level of 5%, so the null hypothesis is rejected. Thus, alternative hypothesis is true. It can be said that the survival of the patient is dependent of the

total charges a patient is charged. The strength of this association is a weak association, which can be stated by the Contingency Coefficient for this test.

$H_0$  = Survival of the patient is independent of the total charges a patient is charged.

$H_a$  = Survival of the patient is dependent of the total charges a patient is charged

Statistic	DF	Value	Prob
Chi-Square	5	201.5866	<.0001
Likelihood Ratio Chi-Square	5	77.6725	<.0001
Mantel-Haenszel Chi-Square	1	160.1396	<.0001
Phi Coefficient		0.2378	
Contingency Coefficient		0.2313	
Cramer's V		0.2378	

**Table.4.27** Statistics for table for Survival vs total charges

From the cross tabulation table (Table 4.28) it can be said that out of patients who spend less than or equal to \$50,000 are 92.29%. 92.98 % survived who paid less than or equal to 10 days, while 79.79% expired in this category. On the other hand, when patients spent \$150,000 -\$200,000 0.22% and 6.25 % women spend \$50001 to \$100000.

		Survival_code		
		Expired	Survived	Total
total_charges Less than or equal to 10	Frequency	150	3141	3291
	Percent	4.21	88.08	92.29
	Row Pct	4.56	95.44	
	Col Pct	79.79	92.98	
50001 to 100000	Frequency	18	205	223
	Percent	0.50	5.75	6.25
	Row Pct	8.07	91.93	
	Col Pct	9.57	6.07	
100001 to 150000	Frequency	7	26	33
	Percent	0.20	0.73	0.93
	Row Pct	21.21	78.79	
	Col Pct	3.72	0.77	
150001 to 200000	Frequency	3	5	8
	Percent	0.08	0.14	0.22
	Row Pct	37.50	62.50	
	Col Pct	1.60	0.15	
200001 to 250000	Frequency	2	1	3
	Percent	0.06	0.03	0.08
	Row Pct	66.67	33.33	
	Col Pct	1.06	0.03	
25000 and more	Frequency	8	0	8
	Percent	0.22	0.00	0.22
	Row Pct	100.00	0.00	
	Col Pct	4.26	0.00	
Total	Frequency	188	3378	3566
	Percent	5.27	94.73	100.00

**Table.4.28** Cross Tabulation: Survival vs total charges

## CHAPTER V

### PREDICTIVE MODELING

This chapter discusses the findings from predictive modeling technique. The first section discusses the results from IBM SPSS Modeler. Three strategies Neural Network- MLP, Neural Network- RBF, and Decision tree are compared using the original dataset the unbalanced vs a balanced dataset using random over sampling techniques. Thus, the study compares Neural Network- MLP using balanced data, Neural Network- RBF using balanced data, and Decision tree using balanced data to Neural Network- MLP using unbalanced data, Network- RBF using unbalanced data and, Decision tree using unbalanced data. These techniques are compared using two different softwares; IBM SPSS modeler and SAS Enterprise Miner 12.3.

#### **5.1 Balanced Data With IBM SPSS Modeler.**

##### **i. Neural Network: Multilayer Perceptron (MLP)**

Table 4.29 shows results from Neural Network, which uses Radial Basis Function (RBF) Procedures. It is a supervised learning technique i.e. they map relationships derived from the data. This rotation estimation technique is a model validation technique which uses the validation dataset in order to minimize the problem of overfitting. 10 cross fold validation method

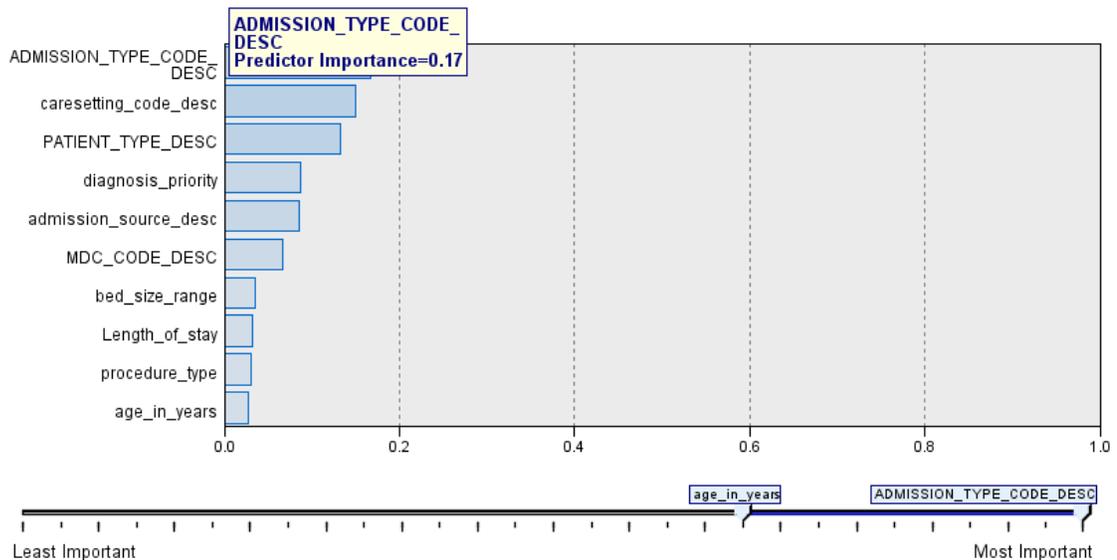
gives an insight on how the model will generalize on an unknown dataset. Steps involved in cross validation are partitioning a sample of data into subsets i.e. in this case 6,574 records are divided into subsets. Then analysis is done on one subset at a time after which validation is done on the other subsets. In order to reduce variability, there are 10 such rounds or subsets on which cross validation is performed using different partitions. The results are then averaged over the subsets. Cross validation when compared to the conventional data partitioning technique is better because the root mean square error that conventional data partitioning technique generates is not a useful estimator of model performance.

Model	Accuracy	Method	Predictors	Model Size (Synapses)	Records
1	99.1%		29	2852	4,410
2	98.6%		29	3060	4,410
3	98.9%		29	3104	4,410
4	99.0%		29	3071	4,410
5	99.1%		29	3386	4,410
6	99.1%		29	3733	4,410
7	98.9%		29	3616	4,410
8	98.8%		29	2852	4,410
9	98.8%		29	3060	4,410
10	99.0%		29	2242	4,410

**Table.4.29** NN- MLP balanced data using IBM SPSS

Thus, the error on the validation dataset does not assess the model performance as well as cross validation. Cross validation thus, averages the measures of prediction error to correct while

training error and obtain accurate estimates of model performance. In this case, 6,574 is divided into roughly 10 equal groups i.e. 4,410 and each of this group has its own model, accuracy, and predictors. As we see from the table above, the number of predictors, i.e. 29 are the same throughout the 10 groups and so are the number of records 4,410. The best model has an accuracy of 99.1% with 29 predictors and the model is tested on 4,410 records. The graph shows (Fig 4.16) the predictor importance of this MLP- Neural Network technique. The most important predictor for predicting the survival of patients is admission type of the patient i.e. urgent, emergency, etc. has a predictor importance of 0.17. The next most important predictor is care-setting type with an importance of 0.15. The next most important predictor is patient admission type like inpatient, outpatient, etc. has a predictor importance of 0.13. The diagnosis priority i.e. Is ovarian cancer a principal diagnosis or secondary diagnosis or so on, has an importance of 0.09. Another parameter with similar predictor importance of 0.09 is admission source type i.e. physician referral, clinical referral etc. Age of the patient and length of stay have an importance of 0.03. Both of these variables are ranked amongst the top ten predictors for survival of ovarian cancer patients using neural network- multilayer perceptron technique.



**Fig.4.16** Predictor importance NN- MLP balanced data using IBM SPSS

Table 4.30 (derived from table Appendices) shows the coincidence matrix. It gives an insight of the true positives, true negatives, false positives, and false negatives for this particular technique. Based on coincidence matrix the accuracy {the ratio between the total numbers of correctly classified cases to the overall number of cases under consideration} for the overall model is 97.71%. Sensitivity {the proportion of positive cases, which are correctly classified i.e. the percentage of patients who expired and classified correctly as expired} 95.47%. Specificity {the proportion of negative cases, which are correctly classified i.e. the percentage of patients who survived and classified correctly as survived} is 100 %.

Where AP: Actual Positive, AN: Actual Negative, PP: Predicted Positive, and PN: Predicted Negative.

	AP	AN	Total
PP	1033	0	1033
PN	49	1061	1110
Total	1082	1061	<b>2143</b>
Accuracy	0.977135	<b>97.71349</b>	
Sensitivity	0.954713	<b>95.47135</b>	
Specificity	1	<b>100</b>	

**Table.4.30** Performance evaluation NN- MLP balanced data using IBM SPSS

ii. **Neural Network: Radial Basis Function (RBF)**

Table 4.31 shows results from Neural Network which uses Radial Basis Function (RBF) procedures. It is a feed forward supervised learning technique. RBF is used along with rotation estimation technique 10 fold cross validation in order to minimize the problem of overfitting. In this case 6,574 records are divided into subsets. Then analysis is done on one subset at a time after which validation is done on the other subsets. In order to reduce variability, there are 10 such rounds or subsets on which cross validation is performed using different partitions. The results are then averaged over the subsets. Here, the 6,574 records of the balanced dataset are

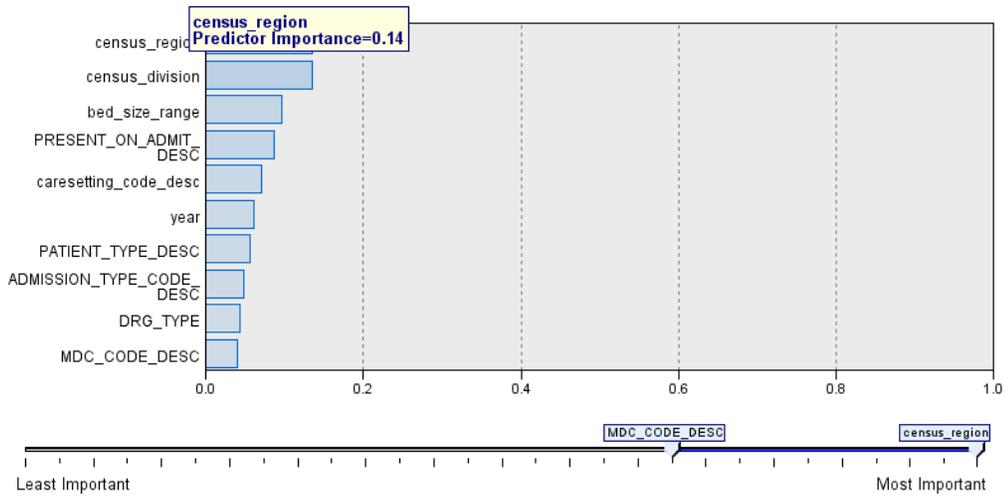
divided into roughly 10 equal groups i.e. 4,410 and each of this group has its own model, accuracy, and predictors. As we see from the table above, the number of predictors, i.e. 29 are the same throughout the 10 groups and so are the number of records 4,410. The best model has an accuracy of 62.5% with 29 predictors and the model is tested on 4,410 records.

Model	Accuracy	Method	Predictors	Model Size (Synapses)	Records
1	62.5%		29	568	4,410
2	76.4%		29	554	4,410
3	64.0%		29	562	4,410
4	78.7%		29	834	4,410
5	62.1%		29	562	4,410
6	60.9%		29	572	4,410
7	61.6%		29	554	4,410
8	76.3%		29	568	4,410
9	59.8%		29	554	4,410
10	69.2%		29	558	4,410

**Table.4.31** NN- RBF balanced data using IBM SPSS

Fig.4.17 is the predictor importance of this RBF- Neural Network technique. The most important predictor for predicting the survival of patients is census region a patient belongs to i.e. South, Midwest, etc. and census division, both having equal predictor importance of 0.14. The next most important predictor with an importance of 0.10 is the bed size category of the hospital.

The variable present on admit has an importance of 0.09, while care-setting type has an importance of 0.07. Using the RBF technique the importance of admission type of the patient has decreased from 0.17 in MLP to a 0.05. In addition, the type of patient admission like inpatient, outpatient, etc. has decreased its predictor importance from 0.13 in MLP to 0.06 in RBF technique. However, the type of DRG and MDC code group both of these variables have an importance of 0.04 and are amongst the top ten predictors for survival of ovarian cancer patients using neural network- radial basis function technique.



**Fig.4.17** Predictor importance NN- RBF balanced data using IBM SPSS

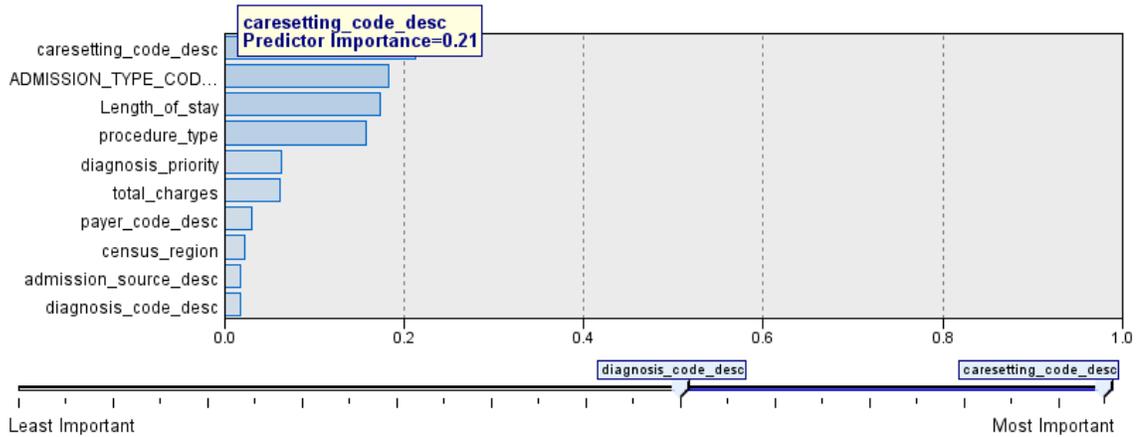
Table 4.32 (derived from table Appendices) shows the coincidence matrix. Its gives an insight of the true positives, true negatives, false positives, and false negatives for this particular technique. The accuracy for the overall model is 67.8%, sensitivity 71.62% and specificity 63.90%.

	AP	AN	Total
PP	775	383	1158
PN	307	678	985
Total	1082	1061	<b>2143</b>
<b>Accuracy</b>	0.678021	<b>67.80215</b>	
<b>Sensitivity</b>	0.716266	<b>71.62662</b>	
<b>Specificity</b>	0.63902	<b>63.90198</b>	

**Table.4.32** Performance evaluation NN- RBF balanced data using IBM SPSS

**i. C5 Decision Tree: 10 fold cross validation**

A C5 decision tree is generated using 10 fold cross validation technique. The depth of the tree is 14, while cross validation mean is 95.8 with a standard error 0.3. The first leaf of the tree is length of stay  $\leq 0$  and length of stay  $> 0$ . The first leaf length of stay  $\leq 0$  has further two leaves total charges  $\leq 23239.50$  and total charges  $> 23239.50$ . Similarly, the tree expands until it has 14 leaf. The predictor importance for this technique is shown in the graph Fig.4.18.



**Fig.4.18** Predictor importance decision trees balanced data using IBM SPSS

Using C5 decision tree the most important variable is care-setting type of the patient, which has an importance of 0.21. The admission type is the second most important predictor of survival with an importance of 0.18. The predictor importance of admission type when compared to the other two techniques is higher for MLP 0.17 and for RBF 0.05. While length of stay which was amongst the top 10 predictors in MLP is the third most important predictor using decision tree with an importance of 0.17. The procedure type i.e. the type of surgery, test etc. conducted on the patient has an importance of 0.16. Census region, which was the first most important predictor in RBF with importance of 0.14, has an importance of 0.02, which is similar of admission source and diagnosis code.

Table 4.33 (derived from table Appendices) calculates the accuracy for the overall model i.e. 96.07%, sensitivity 92.29% and specificity as 100%.

	AP	AN	Total
PP	1018	0	1018
PN	85	1061	1146
Total	1103	1061	<b>2164</b>
<b>Accuracy</b>	0.960721	<b>96.07209</b>	
<b>Sensitivity</b>	0.922937	<b>92.29374</b>	
<b>Specificity</b>	1	<b>100</b>	

**Table.4.33** Performance evaluation decision tree balanced data using IBM SPSS

## 5.2 Unbalanced Data With IBM SPSS Modeler.

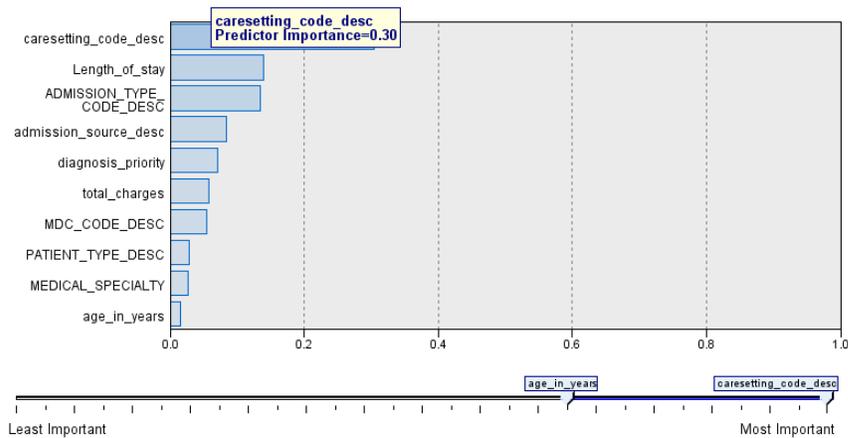
### i. Neural Network: Multilayer Perceptron (MLP)

Table 4.34 shows results from Neural Network, which uses Multilayer Perceptron (MLP) procedures. The rotation estimation technique is used to minimize the problem of overfitting. In this case, 3,566 records are divided into roughly 10 equal groups i.e. 2,359 and each of this group has its own model, accuracy, and predictors. As we see from the table above, the number of predictors, i.e. 29 are the same throughout the 10 groups and so are the number of records 2,359. The best model has an accuracy of 97.6% with 29 predictors and the model is tested on 2,359 records.

Model	Accuracy	Method	Predictors	Model Size (Synapses)	Records
1	97.6%		29	1941	2,359
2	97.9%		29	1146	2,359
3	97.9%		29	1997	2,359
4	97.1%		29	1114	2,359
5	97.6%		29	1432	2,359
6	97.5%		29	1437	2,359
7	97.9%		29	1652	2,359
8	97.1%		29	1754	2,359
9	97.6%		29	1688	2,359
10	97.9%		29	1730	2,359

**Table.4.34** NN- MLP unbalanced data using IBM SPSS

The graph Fig.4.19 shows the predictor importance for unbalance data using MLP- Neural Network technique.



**Fig.4.19** Predictor importance NN- MLP unbalanced data using IBM SPSS

The most important predictor is care-setting type of the patient, which has an importance of 0.30. The length of stay is the second most important predictor of survival with an importance of 0.14 using unbalanced data with MLP technique, length of stay was amongst the top 10 predictors in balanced data using MLP is the third most important predictor for balance data using decision tree with an importance of 0.17. The predictor importance of admission type when compared to the balance data MLP techniques 0.17 and for unbalanced data using MLP it is 0.13. A predictor importance of 0.08 is for predicting the survival of patients given admission source of the patient. The next most important predictor is the diagnosis priority, which has an importance of 0.07. Total charges has an importance of 0.06, while patient type and medical specialty i.e. type of doctor attending the patient has an importance of 0.03 Age of the patient has an importance of 0.02 in this technique.

Table 4.35 (derived from table Appendices) shows the coincidence matrix. The accuracy for the overall model is 94.14%, sensitivity 96.96% and specificity 39.65%.

	AP	AN	Total
PP	1087	35	1122
PN	34	23	57
Total	1121	58	<b>1179</b>
Accuracy	0.941476	<b>94.14758</b>	
Sensitivity	0.96967	<b>96.96699</b>	
Specificity	0.396552	<b>39.65517</b>	

**Table.4.35** Performance evaluation NN- MLP unbalanced data using IBM SPSS

ii. **Neural Network: Radial Basis Function (RBF)**

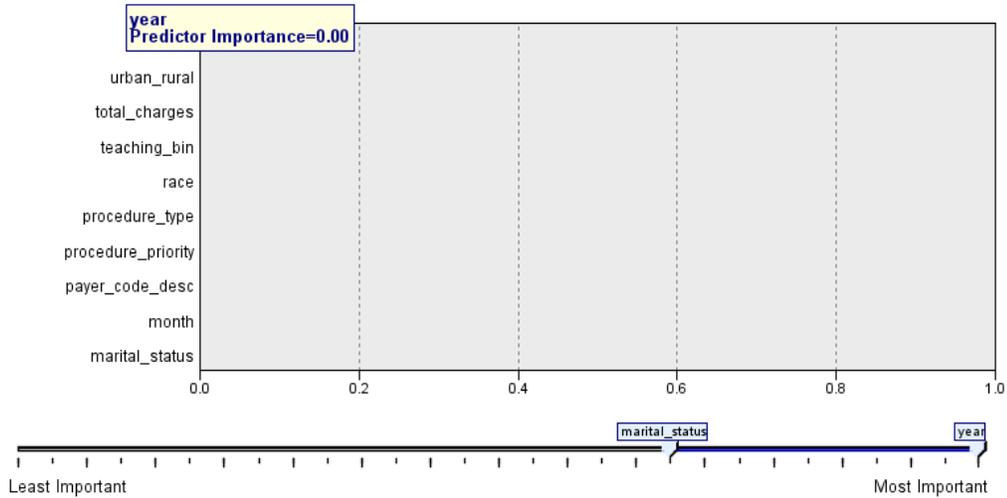
Table 4.36 below shows results from Neural Network which uses Radial Basis Function (RBF) procedures. It is a feed forward supervised learning technique. RBF is used along with rotation estimation technique 10 fold cross validation in order to minimize the problem of overfitting. In

this case, 3,566 records are divided into subsets. These 3,566 records of the unbalanced dataset are divided into roughly 10 equal groups i.e. 2,359 and each of this group has its own model, accuracy, and predictors. As we see from the table above, the number of predictors, i.e. 29 are the same throughout the 10 groups and so are the number of records 2,359. The best model has an accuracy of 94.6% with 29 predictors and the model is tested on 2,359 records.

Model	Accuracy	Method	Predictors	Model Size (Synapses)	Records
1	94.6%		29	552	2,359
2	94.6%		29	570	2,359
3	94.6%		29	568	2,359
4	94.6%		29	554	2,359
5	94.6%		29	570	2,359
6	94.6%		29	572	2,359
7	94.6%		29	548	2,359
8	94.6%		29	582	2,359
9	94.6%		29	560	2,359
10	94.6%		29	574	2,359

**Table.4.36** NN- RBF unbalanced data using IBM SPSS

Fig.4.20 shows the predictor importance of unbalanced data using RBF- Neural Network technique. As we see, the graph has all the predictor importance as 0. The variables that graph shows are the predictor frequent variables.



**Fig.4.20** Predictor importance NN- RBF unbalanced data using IBM SPSS

However, the performance measurement parameters can give a better insight on the model performance. Accuracy for the overall model is 95.08%, sensitivity 100% and specificity 0% (table 4.37).

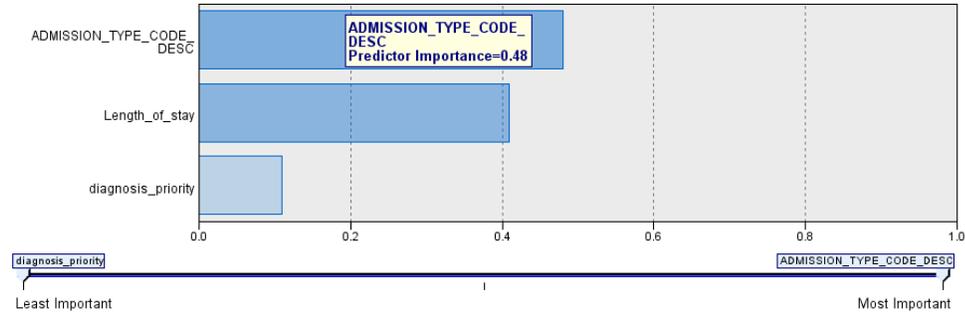
	AP	AN	Total
PP	1121	58	1179
PN	0	0	0
Total	1121	58	<b>1179</b>
<b>Accuracy</b>	0.950806	<b>95.08058</b>	
<b>Sensitivity</b>	1	<b>100</b>	
<b>Specificity</b>	0	<b>0</b>	

**Table.4.37** Performance evaluation NN- RBF unbalanced data using IBM SPSS

**iii. C5 Decision Tree: 10 fold cross validation**

A C5 decision tree is generated using 10 fold cross validation technique. The depth of the tree is 5, while cross validation mean is 94.4 with a standard error 0.2. The first leaf of the tree is length of stay < =6 and length of stay >6. The first leaf length of stay < =0 has further two leaves

diagnosis priority (1 to 3) and diagnosis priority (4 and greater). Similarly, the tree expands until it has 5 leaf. The predictor importance for this technique is shown in the graph below (Fig.4.21).



**Fig.4.21** Predictor importance decision tree unbalanced data using IBM SPSS

Using C5 decision tree the most important variable is admission type of the patient, which has an importance of 0.48. The length of stay is the second most important predictor of survival with an importance of 0.41. While the diagnosis priority has an importance of 0.11. The graph (Fig.4.21) only 3 important variables unlike other techniques where there were 10 predictors for each technique. Table 4.38 is used to calculate the accuracy for the overall model i.e. 94.86%, sensitivity 98.95%, and specificity 16.67%. AP: Actual Positive, AN: Actual Negative, PP: Predicted Positive, and PN: Predicted Negative.

	AP	AN	Total
PP	1135	50	1185
PN	12	10	22
Total	1147	60	<b>1207</b>
<b>Accuracy</b>	0.948633	<b>94.8633</b>	
<b>Sensitivity</b>	0.989538	<b>98.9538</b>	
<b>Specificity</b>	0.166667	<b>16.6667</b>	

**Table.4.38** Performance evaluation decision tree unbalanced data using IBM SPSS

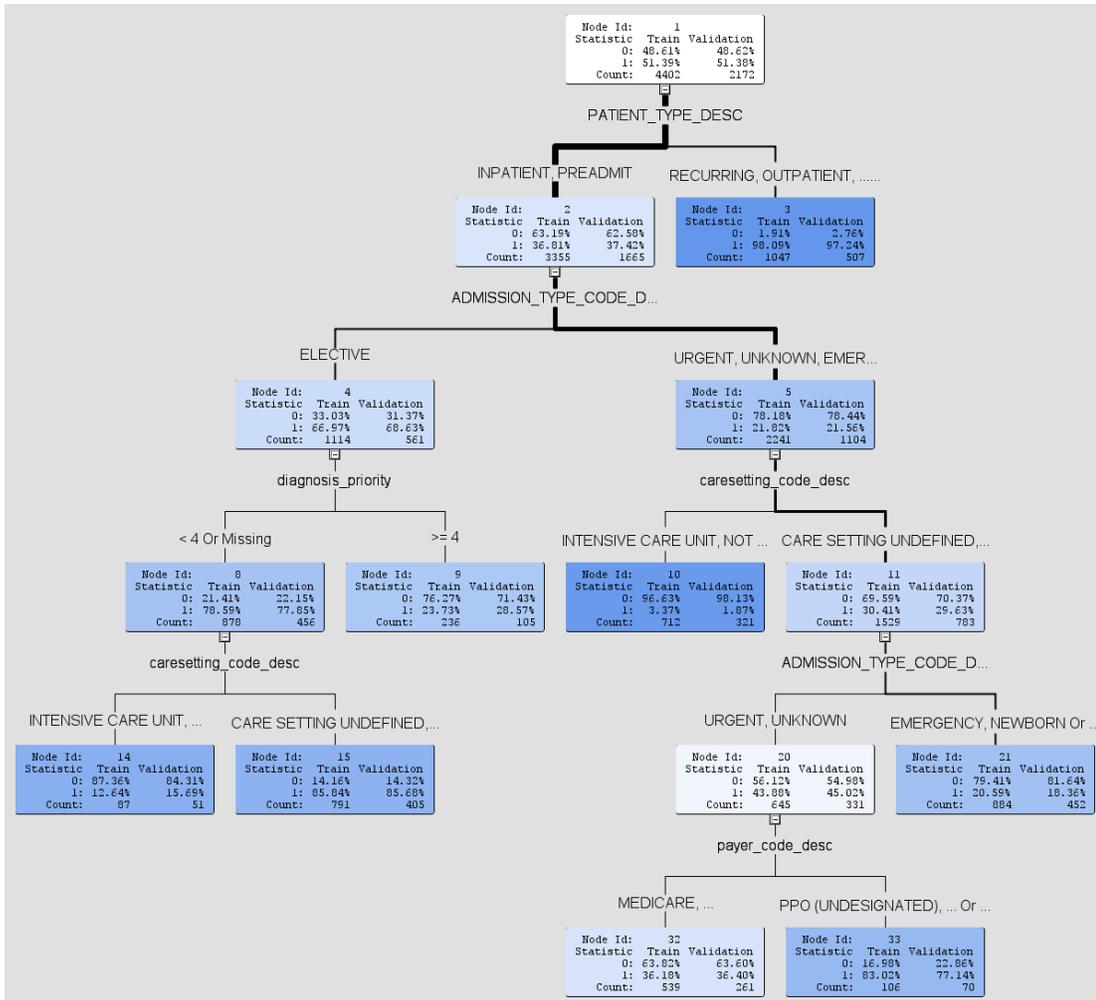
This part of this chapter discusses the predictive modeling technique findings using SAS Enterprise Miner 12.3. Thus, the results from Neural Network- MLP using balanced data, Neural

Network- RBF using balanced data, Decision tree using balanced data , Neural Network- MLP using unbalanced data, Network- RBF using unbalanced data and, Decision tree using unbalanced data are discussed below. Later, the best model is evaluated using performance measurement technique.

### **5.3 Balanced Data With SAS Enterprise Miner 12.3**

#### **i. Neural Network: Multilayer Perceptron (MLP)**

For this study, the inputs for neural network are selected using Stepwise logistic regression. The advantage of this technique is to reduce the number of inputs that the neural network would use for its analysis. Thus, the first step in this strategy is to perform logistic regression and find useful inputs then use those inputs for neural network for this technique the number of inputs are reduced to 16 inputs. Neural networks are complex and difficult to explain. These are some of the reasons why they are considered black box. This study uses decision tree to explain the outputs from neural network since decision tree are simpler to understand and have minimal complexity. The tree (Fig.4.22) shows the results from neural network in a tree format.



**Fig.4.22** Decision tree: NN –MLP balanced data using SAS

Depth of the tree for the decision tree above is 8. The tree shows the first split using the variable patient type. The tree can be explained by an IF then clause. Given that:

PATIENT\_TYPE\_DESC IS ONE OF: RECURRING, OUTPATIENT, OUTPATIENT SURGERY, OTHER SPECIALTY, OBSERVATION / SHORT STAY / 24 HR, EMERGENCY, SERIES, DAY SURGERY, RADIOLOGY, CLINIC, OBSERVATION or MISSING

Then, the predicted probability of survival is 0.98 and patient being expired is 0.02. The other leaf that could be explained as given the diagnosis\_priority >= 4

AND PATIENT\_TYPE\_DESC IS ONE OF: INPATIENT, PREADMIT

AND ADMISSION\_TYPE\_CODE\_DESC IS ONE OF: ELECTIVE then, the predicted probability of survival is 0.24 and patient being expired is 0.76. Similarly, as we go down the tree the probabilities vary (see Appendices: Decision tree rules for balanced dataset using neural network MLP). Table 4.39 below shows the variable importance for this technique.

Variable Name	Label	Number of Splitting Rules	Importance	Validation Importance	Ratio of Validation to Training Importance
PATIENT_TYPE_DESC		1	1.0000	1.0000	1.0000
ADMISSION_TYPE_CODE_DESC		2	0.7574	0.8297	1.0954
caresetting_code_desc		2	0.5086	0.5338	1.0495
diagnosis_priority		1	0.4322	0.3851	0.8910
payer_code_desc		1	0.2547	0.2539	0.9972

**Table.4.39** Predictor importance NN- MLP balanced data using SAS

Table 4.40 can be explained as patient type having an importance of 1 which can be considered as the base variable or reference variable. The next most important variable is admission type, which has an importance of 0.758 when compared to patient type. Care-setting type has an importance of 0.509 i.e. it is 50% less important than patient type. While diagnosis priority has an importance of 0.432 and payer type has an importance of 0.255. The next table describes the statistics for this model.

Statistics Label	Train	Validation
Sum of Frequencies	4402.00	2172.00
Misclassification Rate	0.14	0.14
Maximum Absolute Error	0.98	0.98
Sum of Squared Errors	950.39	477.52
Average Squared Error	0.11	0.11
Root Average Squared Error	0.33	0.33
Divisor for ASE	8804.00	4344.00
Total Degrees of Freedom	4402.00	.

**Table.4.40** Statistics for NN- MLP balanced data using SAS

Misclassification rate for this model is 0.14 in both training as well as validation data. Misclassification occurs when the record or observation belongs to one class, but the model

classifies it as a member of another class. For example, misclassification in this study means, a patient has survived, but the model classifies it as a member of expired group. The sum of squared errors tend to decrease from 950.0 in training dataset to a 477.5 in validation dataset.

Table 4.41 is used to calculate the accuracy for the overall model 86.42%, sensitivity 84.59%, and specificity 88.35%.

	AP	AN	Total
PP	944	123	1067
PN	172	933	1105
Total	1116	1056	<b>2172</b>
<b>Accuracy</b>	0.86418	<b>86.418</b>	
<b>Sensitivity</b>	0.845878	<b>84.5878</b>	
<b>Specificity</b>	0.883523	<b>88.3523</b>	

**Table.4.41** Performance evaluation NN- MLP balanced data using SAS

Where AP: Actual Positive, AN: Actual Negative, PP: Predicted Positive, and PN: Predicted Negative.

**ii. Neural Network: Radial Basis Function (RBF)**

In the input selection step for neural network with radial basis function, there are 16 inputs. The tree (Fig.4.23) shows the results from neural network in a tree format. Depth of the tree for the decision tree above is 8. The tree shows the first split using the variable patient type. The tree can be explained by an IF then clause. Given that:

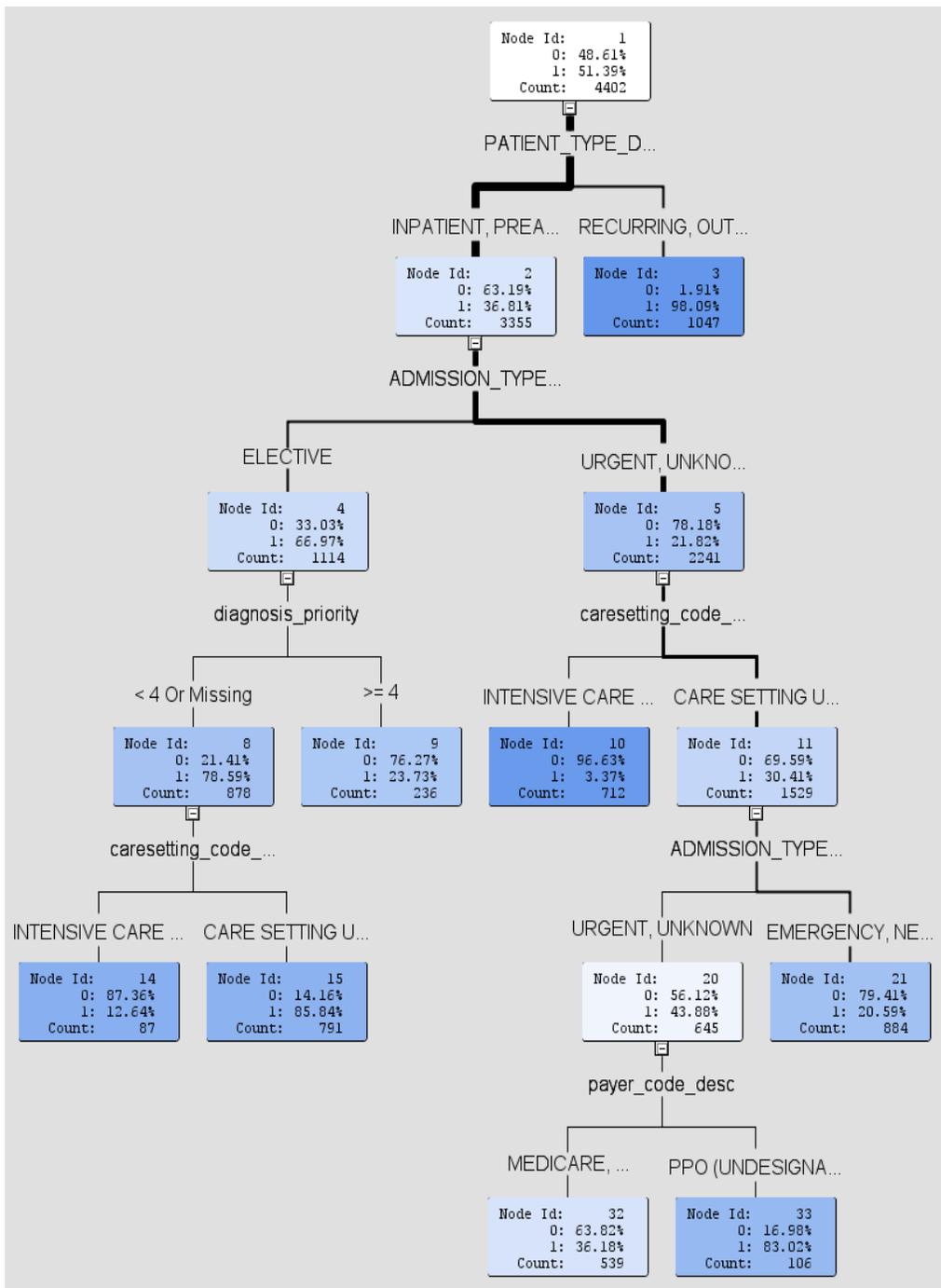
If PATIENT\_TYPE\_DESC IS ONE OF: RECURRING, OUTPATIENT, OUTPATIENT SURGERY, OTHER SPECIALTY, OBSERVATION / SHORT STAY / 24 HR, EMERGENCY, SERIES, DAY SURGERY, RADIOLOGY, CLINIC, OBSERVATION or MISSING

Then, the predicted probability of survival is 0.98 and patient being expired is 0.02.

The other leaf that could be explained as given the

diagnosis\_priority >= 4 AND PATIENT\_TYPE\_DESC IS ONE OF: INPATIENT, PREADMIT

AND ADMISSION\_TYPE\_CODE\_DESC IS ONE OF: ELECTIVE given that both of these and conditions are fulfilled then, the predicted probability of survival is 0.24 and patient being expired is 0.76. Similarly, as we go down the tree the probabilities vary (see Appendices: Decision tree rules for balanced dataset using neural network RBF).



**Fig.4.23** Decision tree: NN- RBF balanced data using SAS

Table 4.42 below shows the variable importance for this technique.

Variable Name	Label	Number of Splitting Rules	Importance	Number of Rules in CV Trees	Relative Importance	Validation Importance	Ratio of Validation to Training Importance
PATIENT_TYPE_DESC		1	1.0000	10	1.0000	1.0000	1.0000
ADMISSION_TYPE_CODE_DESC		2	0.7574	15	0.7354	0.7274	0.9604
caresetting_code_desc		2	0.5086	24	0.5365	0.5154	1.0134
diagnosis_priority		1	0.4322	13	0.4443	0.4265	0.9867
payer_code_desc		1	0.2547	4	0.1600	0.1683	0.6607

**Table.4.42** Predictor importance NN- RBF balanced data using SAS

Table 4.43 can be explained as patient type having an importance of 1, which is the base variable or reference variable. The next most important variable is admission type, which has an importance of 0.758 when compared to patient type. Care-setting type has an importance of 0.509 i.e. it is 50% less important than patient type. While diagnosis priority has an importance of 0.432 and payer type has an importance of 0.255. The next table describes the statistics for this model.

Statistics Label	Train	Validation
Sum of Frequencies	4402.00	2172.00
Misclassification Rate	0.14	0.14
Maximum Absolute Error	0.98	0.98
Sum of Squared Errors	950.39	477.52
Average Squared Error	0.11	0.11
Root Average Squared Error	0.33	0.33
Divisor for ASE	8804.00	4344.00
Total Degrees of Freedom	4402.00	.

**Table.4.43** Statistics for NN- RBF balanced data using SAS

Misclassification rate for this model is 0.14 in both training as well as validation data. While the sum of squared errors tend to decrease from 950.0 in training dataset to a 477.5 in validation dataset. The results from both of the neural network model seems to be very similar in terms of statistics.

Table 4.44 below is used to calculate the accuracy for the overall model 89.25%, sensitivity 83.42%, and specificity 95.44%. AP: Actual Positive, AN: Actual Negative, PP: Predicted Positive, and PN: Predicted Negative.

	AP	AN	Total
PP	931	48	979
PN	185	1008	1193
Total	1116	1056	<b>2172</b>
<b>Accuracy</b>	0.892577	<b>89.2726</b>	
<b>Sensitivity</b>	0.834229	<b>83.4229</b>	
<b>Specificity</b>	0.954416	<b>95.4545</b>	

**Table.4.44** Performance evaluation NN- RBF balanced data using SAS

### iii. Decision Tree

The decision tree is used with cross validation technique to minimize overfitting. The tree (Fig.4.24) is splitting rules for this tree can be explained using the If then clause. Given that

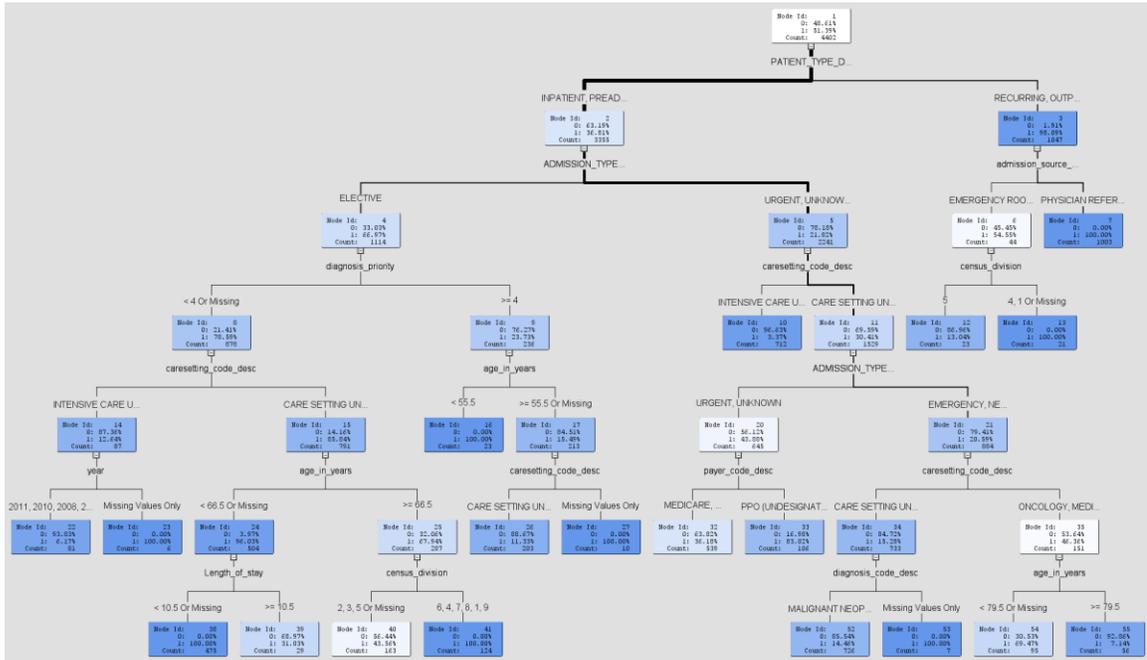
IF admission\_source\_desc IS ONE OF: PHYSICIAN REFERRAL, CLINIC REFERRAL, UNKNOWN or MISSING AND PATIENT\_TYPE\_DESC IS ONE OF: RECURRING, OUTPATIENT, OUTPATIENT SURGERY, OTHER SPECIALTY, OBSERVATION / SHORT STAY / 24 HR, EMERGENCY, SERIES, DAY SURGERY, RADIOLOGY, CLINIC, OBSERVATION or MISSING

then, the predicted probability of survival is 1 and patient being expired is 0.00.

The other leaf that could be explained as given the

if caresetting\_code\_desc IS ONE OF: INTENSIVE CARE UNIT, NOT MAPPED, INTENSIVE CARE UNIT - MEDICAL, CORONARY CARE UNIT, INTENSIVE CARE UNIT - SURGICAL, UROLOGY AND PATIENT\_TYPE\_DESC IS ONE OF: INPATIENT, PREADMIT

AND ADMISSION\_TYPE\_CODE\_DESC IS ONE OF: URGENT, UNKNOWN, EMERGENCY, NEWBORN or MISSING then, the predicted probability of survival is 0.03 and patient being expired is 0.97 (see Appendices: Decision tree rules for balanced dataset using SAS).



**Fig.4.24** Decision tree for balanced data using SAS

Table 4.45 below shows the variable importance for this technique.

Variable Name	Label	Number of Splitting Rules	Importance	Validation Importance	Ratio of Validation to Training Importance
PATIENT_TYPE_DESC		1	1.0000	1.0000	1.0000
ADMISSION_TYPE_CODE_DESC		2	0.7574	0.8297	1.0954
caresetting_code_desc		5	0.5748	0.5904	1.0272
diagnosis_priority		1	0.4322	0.3851	0.8910
age_in_years		3	0.3785	0.3605	0.9523
census_division		2	0.3203	0.3028	0.9454
Length_of_stay		2	0.2588	0.2857	1.1041
payer_code_desc		1	0.2547	0.2539	0.9972
month		2	0.1939	0.1918	0.9892
admission_source_desc		1	0.1705	0.2405	1.4107
year		1	0.1281	0.1266	0.9880

**Table.4.45** Predictor importance balanced data: Decision tree

Table 4.46 can be explained as patient type having an importance of 1, which can be considered as the base variable or reference variable. The next most important variable is admission type, which has an importance of 0.758 when compared to patient type. Care-setting type has an importance of 0.57. While diagnosis priority has an importance of 0.432 and age has an importance of 0.38. The other variables that hold importance from 0.32 to 0.12 are census division where a person belongs, length of stay, payer type, month of admit, admission source and the year of admit. The next table describes the statistics for this model.

Statistics Label	Train	Validation
Sum of Frequencies	4402.00	2172.00
Misclassification Rate	0.11	0.12
Maximum Absolute Error	0.99	0.99
Sum of Squared Errors	670.15	344.90
Average Squared Error	0.08	0.08
Root Average Squared Error	0.28	0.28
Divisor for ASE	8804.00	4344.00
Total Degrees of Freedom	4402.00	.

**Table.4.46** Statistics for decision tree balanced data using SAS

Misclassification rate for this model is 0.12 in validation data and 0.11 in training data. While the sum of squared errors tend to decrease from 670.15 in training dataset to 344.90 in validation dataset. Table 4.47 is used to calculate the accuracy for the overall model 87.93%, sensitivity 79.9%, and specificity 96.40%. AP: Actual Positive, AN: Actual Negative, PP: Predicted Positive, and PN: Predicted Negative.

	AP	AN	Total
PP	892	38	930
PN	224	1018	1242
Total	1116	1056	<b>2172</b>
Accuracy	0.879374	<b>87.9374</b>	
Sensitivity	0.799283	<b>79.9283</b>	
Specificity	0.964015	<b>96.4015</b>	

**Table.4.47** Performance evaluation decision tree balanced data using SAS

## 5.4 Unbalanced Data With SAS Enterprise Miner 12.3

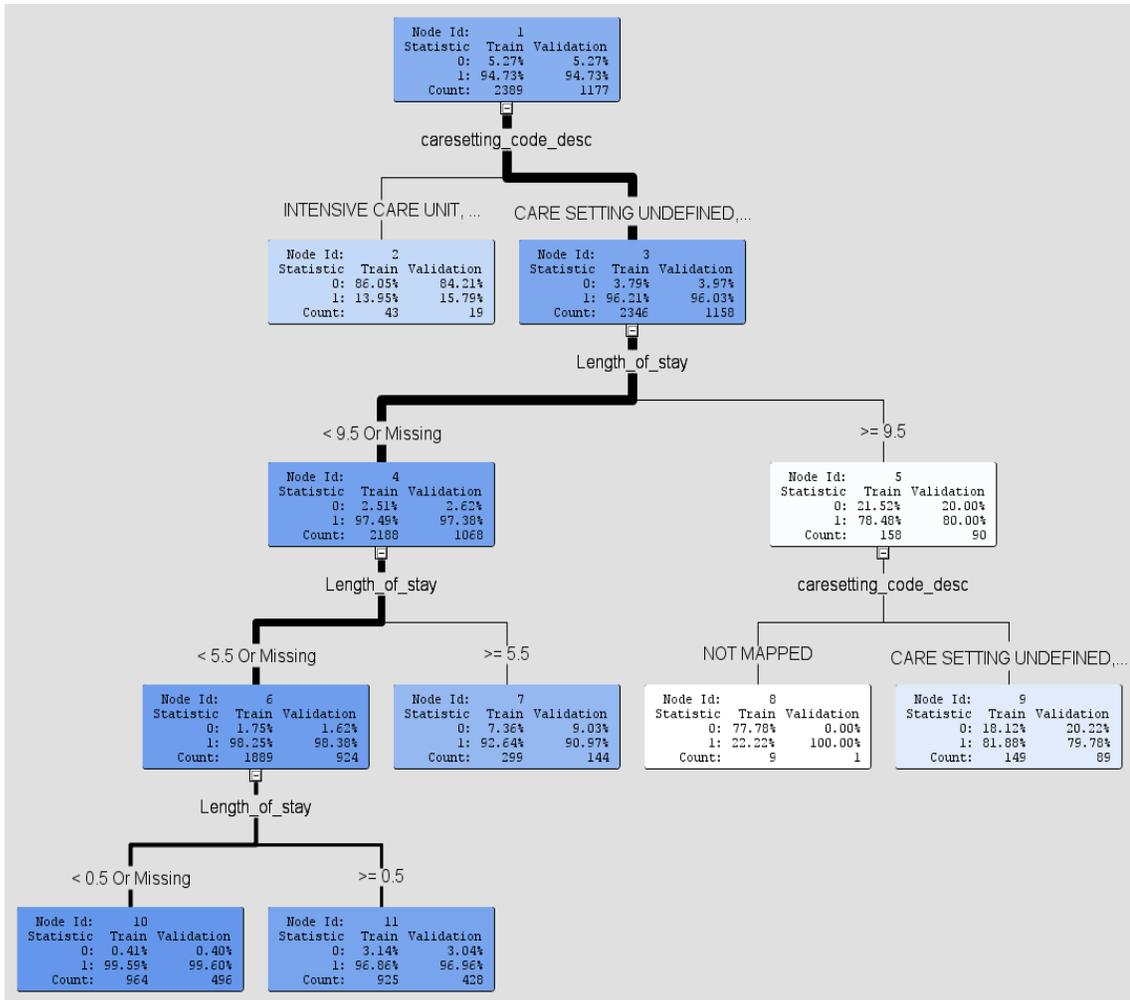
### i. Neural Network: Multilayer Perceptron (MLP)

The model selects two inputs after stepwise input selection from logistic regression. Decision tree is used to explain the outputs from neural network since decision tree are simpler to understand and have minimal complexity. The tree (Fig.4.25) show the results from neural network in a tree format. Depth of the tree for the decision tree above is 5. The tree shows the first split using the variable patient type. The tree can be explained by an IF then clause. Given that:

If caresetting\_code\_desc IS ONE OF: INTENSIVE CARE UNIT, INTENSIVE CARE UNIT - MEDICAL, INTENSIVE CARE UNIT - SURGICAL, CORONARY CARE UNIT.

Then, the predicted probability of survival is 0.14 and patient being expired is 0.86.

The other leaf that could be explained as given if caresetting\_code\_desc IS ONE OF: CARE SETTING UNDEFINED, ONCOLOGY, GYNECOLOGY, MULTISPECIALTY UNIT, NOT MAPPED, OBSTETRICS, MEDICAL/SURGICAL, SURGERY, OUTPATIENT CLINIC, AMBULATORY SURGERY, AMBULAT AND Length\_of\_stay < 9.5 AND Length\_of\_stay >= 5.5. Then the probability of survival is 0.93 and patient being expired is 0.07. Similarly, as we go down the tree the probabilities vary (see Appendices: Decision tree rules for unbalanced dataset using neural network MLP).



**Fig.4.25** Decision tree: NN- MLP unbalanced data using SAS

Table 4.48 below shows care-setting type having an importance of 1 which is the base variable or reference variable. The next most important variable is length of stay, which has an importance of 0.453 when compared to care-setting type.

Variable Name	Label	Number of Splitting Rules	Importance	Validation Importance	Ratio of Validation to Training Importance
caresetting_code_desc		2	1.0000	1.0000	1.0000
Length_of_stay		3	0.4531	0.5232	1.1547

**Table.4.48** Predictor importance NN- MLP unbalanced data using SAS

Statistics Label	Train	Validation
Sum of Frequencies	2389.00	1177.00
Misclassification Rate	0.04	0.04
Maximum Absolute Error	1.00	1.00
Sum of Squared Errors	162.56	88.00
Average Squared Error	0.03	0.04
Root Average Squared Error	0.18	0.19
Divisor for ASE	4778.00	2354.00
Total Degrees of Freedom	2389.00	.

**Table.4.49** Statistics for NN- MLP unbalanced data using SAS

Table 4.49 shows the statistics for this model. Misclassification rate for this model is 0.04 in both training as well as validation data. The sum of squared errors tend to decrease from 162.56 in training dataset to 88.0 in validation dataset. The table below shows the confusion matrix, for training and validation.

The accuracy for the overall model 96.18%, sensitivity 99.73%, and specificity 32.26% (table 4.50).

	AP	AN	Total
PP	1112	42	1154
PN	3	20	23
Total	1115	62	<b>1177</b>
Accuracy	0.961767	<b>96.1767</b>	
Sensitivity	0.997309	<b>99.7309</b>	
Specificity	0.322581	<b>32.2581</b>	

**Table.4.50** Performance evaluation NN- MLP unbalanced data using SAS

Where AP: Actual Positive, AN: Actual Negative, PP: Predicted Positive, and PN: Predicted Negative.

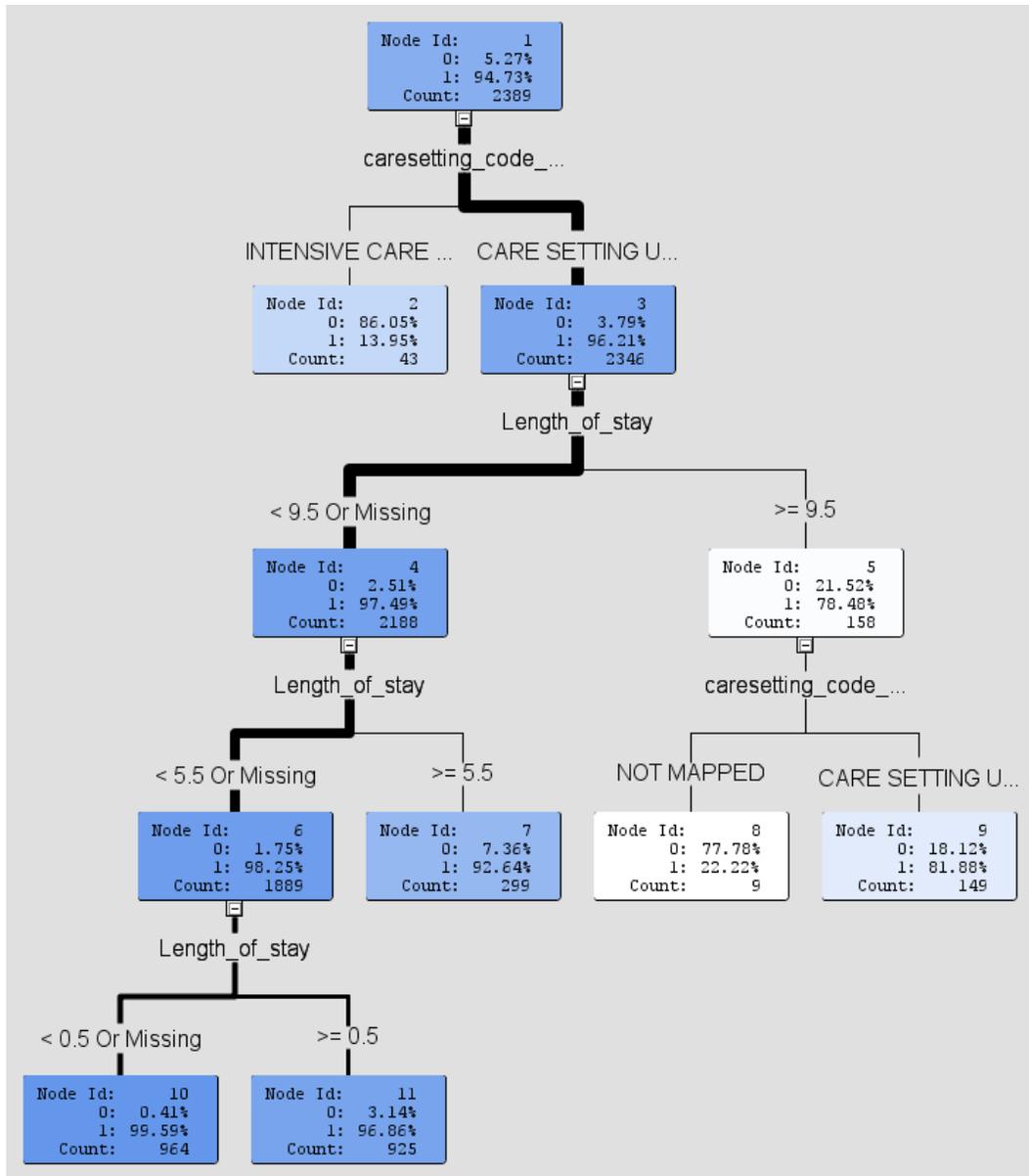
**ii. Neural Network: Radial Basis Function (RBF)**

In the input selection step for neural network with radial basis function, there are 2 inputs. The tree (fig.2.26) below show the results from neural network in a tree format. Depth of the tree for

the decision tree above is 5. The tree shows the first split using the variable patient type. The tree can be explained by an IF then clause. Given that: caresetting\_code\_desc IS ONE OF: INTENSIVE CARE UNIT, INTENSIVE CARE UNIT - MEDICAL, INTENSIVE CARE UNIT - SURGICAL, CORONARY CARE UNIT. Then, the predicted probability of survival is 0.14 and patient being expired is 0.86. The other leaf that could be explained as given

If caresetting\_code\_desc IS ONE OF: CARE SETTING UNDEFINED, ONCOLOGY, GYNECOLOGY, MULTISPECIALTY UNIT, NOT MAPPED, OBSTETRICS, MEDICAL/SURGICAL, SURGERY, OUTPATIENT CLINIC, AMBULATORY SURGERY, AMBULAT AND Length\_of\_stay < 9.5 AND Length\_of\_stay >= 5.5.

Then the probability of survival is 0.93 and patient being expired is 0.07 (see Appendices: Decision tree rules for unbalanced dataset using neural network RBF).



**Fig.2.26** Decision tree: NN- RBF unbalanced data using SAS

Table 4.51 below shows care-setting type having an importance of 1. The next most important variable is length of stay, which has an importance of 0.453 when compared to care-setting type.

Variable Name	Label	Number of Splitting Rules	Importance	Number of Rules in CV Trees	Relative Importance	Validation Importance	Ratio of Validation to Training Importance
caresetting_code_desc		2	1.0000	15	1.0000	1.0000	1.0000
Length_of_stay		3	0.4531	9	0.3919	0.3780	0.8343

**Table.4.51** Predictor importance NN- RBF unbalanced data using SAS

Statistics Label	Train	Validation
Sum of Frequencies	2389.00	1177.00
Misclassification Rate	0.04	0.04
Maximum Absolute Error	1.00	1.00
Sum of Squared Errors	162.56	88.00
Average Squared Error	0.03	0.04
Root Average Squared Error	0.18	0.19
Divisor for ASE	4778.00	2354.00
Total Degrees of Freedom	2389.00	.

**Table.4.52** Statistics for NN- RBF unbalanced data using SAS

Misclassification rate for this model (Table 4.52) is 0.04 in both training as well as validation data. The sum of squared errors tend to decrease from 162.56 in training dataset to 88.0 in validation dataset.

The overall accuracy of this model is 96.17%, sensitivity 99.82%, and specificity 30.65% (Table 4.53). AP: Actual Positive, AN: Actual Negative, PP: Predicted Positive, and PN: Predicted Negative.

	AP	AN	Total
PP	1113	43	1156
PN	2	19	21
Total	1115	62	<b>1177</b>
Accuracy	0.961767	<b>96.1767</b>	
Sensitivity	0.998206	<b>99.8206</b>	
Specificity	0.306452	<b>30.6452</b>	

**Table.4.53** Performance evaluation NN- RBF unbalanced data using SAS

### iii. Decision Tree

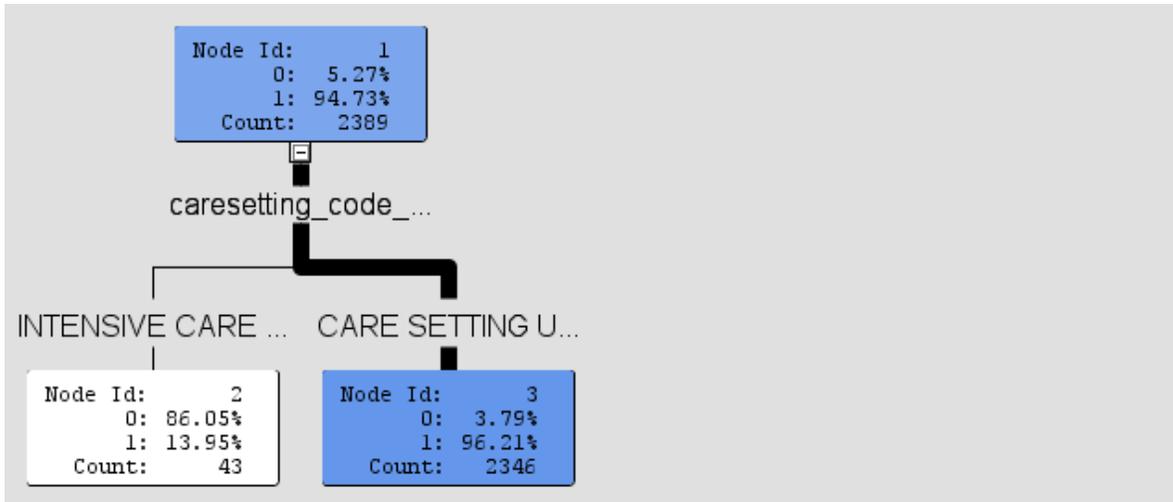
The decision tree is used with cross validation technique to minimize overfitting. The tree (Fig.4.27) is splitting rules and it can be explained using the If then clause. Given that

IF caresetting\_code\_desc IS ONE OF: INTENSIVE CARE UNIT, INTENSIVE CARE UNIT - MEDICAL, INTENSIVE CARE UNIT - SURGICAL, CORONARY CARE UNIT.

Then, the predicted probability of survival is 0.14 and patient being expired is 0.86.

The other leaf that could be explained as given the

If caresetting\_code\_desc IS ONE OF: CARE SETTING UNDEFINED, ONCOLOGY, GYNECOLOGY, MULTISPECIALTY UNIT, NOT MAPPED, OBSTETRICS, MEDICAL/SURGICAL, SURGERY, OUTPATIENT CLINIC, AMBULATORY SURGERY, AMBULAT then, the predicted probability of survival is 0.96 and patient being expired is 0.04 (see Appendices: Decision tree rules for unbalanced dataset using SAS).



**Fig.4.27** Decision tree unbalanced data using SAS

Table 4.54 below shows the variable importance for this technique. The model shows only one important variable having an importance of 1.

Variable Name	Label	Number of Splitting Rules	Importance	Validation Importance	Ratio of Validation to Training Importance
caresetting_code_desc		1	1.0000	1.0000	1.0000

**Table.4.54** Predictor importance Decision tree unbalanced data using SAS

Table 4.55 describes the statistics for this model. Misclassification rate for this model is 0.04 in both training data and validation data. While the sum of squared errors tend to decrease from 181.57 in training dataset to 93.42 in validation dataset.

Statistics Label	Train	Validation
Sum of Frequencies	2389.00	1177.00
Misclassification Rate	0.04	0.04
Maximum Absolute Error	0.96	0.96
Sum of Squared Errors	181.57	93.42
Average Squared Error	0.04	0.04
Root Average Squared Error	0.19	0.20
Divisor for ASE	4778.00	2354.00
Total Degrees of Freedom	2389.00	.

**Table.4.55** Statistics for decision tree unbalanced data using SAS

Accuracy for the overall model is 95.83%, sensitivity 99.73%, and specificity 25.81% (table 4.56). AP: Actual Positive, AN: Actual Negative, PP: Predicted Positive, and PN: Predicted Negative.

	AP	AN	Total
PP	1112	46	1158
PN	3	16	19
Total	1115	62	<b>1177</b>
<b>Accuracy</b>	0.958369	<b>95.8369</b>	
<b>Sensitivity</b>	0.997309	<b>99.7309</b>	
<b>Specificity</b>	0.258065	<b>25.8065</b>	

**Table.4.56** Performance evaluation decision tree unbalanced data using SAS

The last step is to evaluate which technique can best predict survival of ovarian cancer patients. To determine the best model to using balanced technique used by IBM SPSS modeler, using balanced technique used by SAS, using unbalanced technique used by IBM SPSS modeler, and using unbalanced technique used by SAS. The best model using SAS (see Appendices: Model Comparison: Balanced Data) and balanced technique is neural network radial basis

function, the accuracy of the model is 89.27 % which, is followed by Decision tree with an accuracy of 87.94% and the last with an accuracy of 86.42% is neural network multilayer perceptron (Table 4.57). The best model using IBM SPSS modeler and balanced technique is neural network multilayer perceptron, the accuracy of the model is 97.71 % which, is followed by Decision tree with an accuracy of 96.07% and the last with an accuracy of 67.80% is neural network radial basis function. So the results from both the software are compared it is found that **IBM SPSS modeler generates a model with better accuracy 97.71 % and cross validated results for balanced data.**

SAS Output					IBM SPSS Modeler				
<b>Balanced</b>					<b>Balanced</b>				
<b>Neural Net-MLP</b>		AP	AN	Total	<b>Neural Net-MLP</b>		AP	AN	Total
	PP	944	123	1067		PP	1033	0	1033
	PN	172	933	1105		PN	49	1061	1110
	Total	1116	1056	2172		Total	1082	1061	2143
	<b>Accuracy</b>	0.86418	<b>86.418</b>			<b>Accuracy</b>	0.97713	<b>97.7135</b>	
	<b>Sensitivity</b>	0.84588	<b>84.5878</b>			<b>Sensitivity</b>	0.95471	<b>95.4713</b>	
	<b>Specificity</b>	0.88352	<b>88.3523</b>			<b>Specificity</b>	1	<b>100</b>	
<b>Neural Net-RBF</b>		AP	AN	Total	<b>Neural Net-RBF</b>		AP	AN	Total
(normalized radial	PP	931	48	979		PP	775	383	1158
equal width	PN	185	1008	1193		PN	307	678	985
and height )	Total	1116	1056	2172		Total	1082	1061	2143
	<b>Accuracy</b>	0.89273	<b>89.2726</b>			<b>Accuracy</b>	0.67802	<b>67.8021</b>	
	<b>Sensitivity</b>	0.83423	<b>83.4229</b>			<b>Sensitivity</b>	0.71627	<b>71.6266</b>	
	<b>Specificity</b>	0.95455	<b>95.4545</b>			<b>Specificity</b>	0.63902	<b>63.902</b>	
<b>Decision Tree</b>		AP	AN	Total	<b>Decision Tree</b>		AP	AN	Total
	PP	892	38	930		PP	1018	0	1018
	PN	224	1018	1242		PN	85	1061	1146
	Total	1116	1056	2172		Total	1103	1061	2164
	<b>Accuracy</b>	0.87937	<b>87.9374</b>			<b>Accuracy</b>	0.96072	<b>96.0721</b>	
	<b>Sensitivity</b>	0.79928	<b>79.9283</b>			<b>Sensitivity</b>	0.92294	<b>92.2937</b>	
	<b>Specificity</b>	0.96402	<b>96.4015</b>			<b>Specificity</b>	1	<b>100</b>	

**Table.4.57** Overall performance evaluation: Balanced data

On comparing unbalanced technique used by IBM SPSS modeler, and using unbalanced technique used by SAS (Table 4.58). The best model using SAS (see Appendices: Model Comparison: Unbalanced Data) and unbalanced technique is neural network radial basis function, the accuracy of the model is 96.17 % which, is followed by neural network multilayer perceptron

with an accuracy of 96.18% and the last decision tree with an accuracy of 95.84%. The best model using IBM SPSS modeler and balanced technique is neural network radial basis function, the accuracy of the model is 95.08 % which, is followed by Decision tree with an accuracy of 94.86% and the last with an accuracy of 94.15% is neural network multilayer perceptron. So the results from both the software are compared it is found that **SAS generates a model with better accuracy 96.17 % for unbalanced data.**

IBM SPSS Modeler					SAS Output				
UnBalanced					UnBalanced				
<b>Neural Net-MLP</b>					<b>Neural Net-MLP</b>				
	AP	AN	Total		AP	AN	Total		
PP	1087	35	1122	PP	1112	42	1154		
PN	34	23	57	PN	3	20	23		
Total	1121	58	1179	Total	1115	62	1177		
	<b>Accuracy</b>	0.94148	<b>94.1476</b>		<b>Accuracy</b>	0.96177	<b>96.1767</b>		
	<b>Sensitivity</b>	0.96967	<b>96.967</b>		<b>Sensitivity</b>	0.99731	<b>99.7309</b>		
	<b>Specificity</b>	0.39655	<b>39.6552</b>		<b>Specificity</b>	0.32258	<b>32.2581</b>		
<b>Neural Net-RBF</b>					<b>Neural Net-RBF</b>				
	AP	AN	Total		AP	AN	Total		
PP	1121	58	1179	(normalized radial PP	1113	43	1156		
PN	0	0	0	equal width PN	2	19	21		
Total	1121	58	1179	and height ) Total	1115	62	1177		
	<b>Accuracy</b>	0.95081	<b>95.0806</b>		<b>Accuracy</b>	0.96177	<b>96.1767</b>		
	<b>Sensitivity</b>	1	<b>100</b>		<b>Sensitivity</b>	0.99821	<b>99.8206</b>		
	<b>Specificity</b>	0	<b>0</b>		<b>Specificity</b>	0.30645	<b>30.6452</b>		
<b>Decision Tree</b>					<b>Decision Tree</b>				
	AP	AN	Total		AP	AN	Total		
PP	1135	50	1185	PP	1112	46	1158		
PN	12	10	22	PN	3	16	19		
Total	1147	60	1207	Total	1115	62	1177		
	<b>Accuracy</b>	0.94863	<b>94.8633</b>		<b>Accuracy</b>	0.95837	<b>95.8369</b>		
	<b>Sensitivity</b>	0.98954	<b>98.9538</b>		<b>Sensitivity</b>	0.99731	<b>99.7309</b>		
	<b>Specificity</b>	0.16667	<b>16.6667</b>		<b>Specificity</b>	0.25806	<b>25.8065</b>		

**Table.4.58** Overall performance evaluation: Unbalanced data

When both the results of the best models are compared, it is found when an unbalanced data is balanced using random over sampling, **IBM SPSS modeler tends to generate better results with 97.71% accuracy, and if an unbalanced data is, used SAS tends to perform better with 96.18% accuracy.**

## CHAPTER VI

### CONCLUSION

This study applies machine learning techniques for predicting ovarian cancer survivability. Specifically, the study uses three popular data mining techniques: Neural network- MLP, Neural network-RBF and Decision trees. It also provides an insight on the performance of these machine learning techniques using a balanced dataset and an imbalanced dataset. The reason behind using these machine learning techniques are they have been used in cancer detection and diagnosis for nearly 20 years (Simes, 1985). In a real world scenario, most of the cases are imbalanced. Even the dataset used in this study was an imbalanced one i.e. the classification categories, were not in equal representation i.e. the expired class had only 188 records while survived class had 3,378 patient records. Machine learning techniques tend to bias the prediction and are thus, a poor representation of the minority class. Therefore, it is important to investigate how balanced data and imbalanced data perform. The initial number of patient records was 46,792 with 80 factors from the Cerner database.

After cleansing and transformation, the prediction models are generated with 3,566 patient records and 47 variables. This study defines “survival” of ovarian cancer patients as patients who

have been completely cured and discharged home. The “discharged home” classifier was chosen to mark a completely recovered patient because all other classifiers point to a patient who has not fully recovered. This survival was then coded as binary categorical survival variable to represent the survival with a value of “1” and non-survival of value “0”. 10-fold cross validation is used in all the techniques to minimize the overfitting of models. Cross validation divides the dataset into 10 mutually exclusive folds using stratified sampling technique. 9 of 10 folds are used for training and the 10<sup>th</sup> is used for testing. Since it is 10 fold validation, the process is repeated 10 times. In this case each of record is once used as part training and once testing. The performance of all the 10 models is averaged based on the accuracy. This process is repeated for both balanced and unbalanced data for all the three techniques to get an unbiased prediction performance.

Based on the descriptive statistics this study finds similar patterns that are observed in studies conducted on SEER cancer data. The patterns shown by the age of the patient in this study is that the average age of a patient is 60 years. Median age of ovarian cancer patients is 63 years. Married women show a higher rate of ovarian cancer than single, widowed or divorced women do. Based on the race/ ethnicity, Caucasian women show higher chances of having ovarian cancer. Most patients visit a facility that is located in urban area than the rural area. Most of these facilities that are in urban areas belong to the South region, Northeast region, Midwest region and a few in West region. These four census region are further divided into census division where “6” South (DE, DC, FL, GA, MD, NC, SC, VA, WV) has the highest concentration of ovarian cancer patients i.e. 751. Census Division “2” has the second highest concentration of 513 patients in Middle Atlantic (NJ, NY, and PA). The average length of stay for a patient in the hospital is 3.5 days. There were 38.16 % Medicare patients, while 14.55 % Blue Cross/ Blue Shield patients and 7.12% Medicaid patients.

The aggregated results indicate that balanced technique using neural network multilayer perceptron in IBM SPSS modeler performed the best with a classification accuracy of 97.71%

which is better than any other model compared in the study. The second best is using unbalanced data on neural network radial basis function with a classification accuracy of 96.18%. The neural network with radial basis function comes out as the worst with a classification accuracy of 67.80% even with a balanced dataset. This signifies given a set of parameters used in the study like: admission source, race of the patient, census division and so on the neural network using multilayer perceptron will predict the outcome of survival of the patient with 97.71% accuracy. The results from this study clearly points out the potential of neural networks classification technique. The advantages of using ANN is they usually perform better than the other models due to their complex structures that automatically approximates any non-linear mathematical function. They are not sensitive to unusual values (outliers) in the data and thus provide better performance.

This study predicts ovarian cancer survival using two leading software used for predictive modeling SAS Enterprise Miner 9.4 and IBM SPSS 15.0. It is concluded that SAS is very efficient source for data management, robust, requires users to have in depth knowledge of programming and statistics, generates in depth results at once. While SPSS is more user friendly, powerful in graphics, and does not require in depth knowledge about programming. Both of the packages have its own strengths and weaknesses. While making decision on what package works the best the answer depends on various factors like resources available, cost of the packages, knowledge of the user. Therefore, if proper resources are available utilizing both or using mixed models depending on the nature of research is recommended.

This study also investigates how the predictive modeling techniques perform when unbalanced and a balanced dataset is used. From table 4.57 and table 4.58 it is found that prediction accuracy of unbalanced data is comparatively low. The classification techniques show poor performance while handling an unbalanced data and the results are biased towards the majority class. The performance of the models is better when the predictive models have a balanced dataset. The

results show with random over sampling the prediction accuracy is best with 97.71%, AUC (Area under curve) is sensitivity 0.95 and specificity 1 for neural network MLP. Thus, the best models are generated when both the classes are equally represented. To avoid overfitting and increase the performance of the classification technique  $k$ -fold cross validation can be used along with random over sampling technique. In conclusion, a balanced dataset when used along with  $k$ -fold cross validation to generates best models.

In addition to the prediction model, this study also found important factors in order to have a better insight into the relative contribution of the variables to predict survivability. Analysis indicates the top 15 variables of importance are:

1. Admission type.
2. Care-setting type
3. Patient type.
4. Diagnosis priority
5. Length of stay
6. Bed size category of hospital
7. Admission source
8. Procedure type
9. Census region
10. Census division
11. Payer type
12. MDC code
13. Age in years
14. Year of admit
15. Drg type.

Why these factors are more important predictors than the other is a question that can only be answered by medical professional and further clinical studies. This study tries to help find such patterns that might be useful in predicting survival and thus not aiming at replacing the valuable experience of the medical professionals.

A noteworthy strength of this study is that not only does it provide a ranking to the prediction models but also variable importance from these techniques. This will help decision makers understand what variables are the most important in predicting survival given other features like race, payer, admission type and so on. Although data mining methods are useful in pattern recognition, help from medical professionals will always provide better depth to the study. These medical professionals can use their years of experience to evaluate the patterns found in the study and thus categorize these patterns into actionable, logical patterns.

Further, scope of the research can be to look into other types of cancer if they influence or have correlation with ovarian cancer. Second, are there any medication that influence the survival of ovarian cancer patients? Third, further analysis can be conducted on generalizing the software package to be used for analyzing a particular set of data.

## REFERENCES

- Bast Jr, R. C., Urban, N., Shridhar, V., Smith, D., Zhang, Z., Skates, S., . . . Mills, G. (2002). Early detection of ovarian cancer: promise and reality *Ovarian Cancer* (pp. 61-97): Springer.
- Bellaachia, A., & Guven, E. (2006). Predicting breast cancer survivability using data mining techniques. *Age*, 58(13), 10-110.
- Cancer.Net. (2015). from <http://www.cancer.net/cancer-types/ovarian-cancer/diagnosis>.
- Casagrande, J., Pike, M., Ross, R., Louie, E., Roy, S., & Henderson, B. (1979). " Incessant ovulation" and ovarian cancer. *The Lancet*, 314(8135), 170-173.
- Centers for Disease Control and Prevention. (2015). Classification of Diseases, Functioning, and Disability. from <http://www.cdc.gov/nchs/icd/icd9cm.htm>.
- Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16(1), 321-357.
- Cios, K., & Moore, W. G. (2002). Uniqueness of medical data mining. *Artificial Intelligence in Medicine*, 26(1-2), 1-24.
- Eisenhauer, E. L., Tew, W. P., Levine, D. A., Lichtman, S. M., Brown, C. L., Aghajanian, C., . . . Chi, D. S. (2007). Response and outcomes in elderly patients with stages IIIc–IV ovarian cancer receiving platinum–taxane chemotherapy. *Gynecologic Oncology*, 106(2), 381-387.
- Ford, D., Easton, D. F., Bishop, P. D. T., Narod, S. A., & Goldgar, D. E. (1994). Risks of cancer in BRCA1-mutation carriers. *The Lancet*, 343(8899), 692-695.
- Frawley, W., Piatetsky-Shapiro, G., & Matheus, C. (1992). Knowledge Discovery in Databases: An Overview. *AI Magazine*, 13.

<http://www.ovarydisease.com/p/ovarian-cancer.html>.

<http://www.sv-europe.com/crisp-dm-methodology/>.

Huang, M.-J., Chen, M.-Y., & Lee, S.-C. (2007). Integrating data mining with case-based reasoning for chronic diseases prognosis and diagnosis. *Expert Systems with Applications*, 32(3), 856-867.

Kleinbaum, D. G., & Klein, M. (2012). *Survival Analysis: A Self-Learning Text* (Third ed.): Springer .

Kuk, C., Kulasingam, V., Gunawardana, C. G., Smith, C. R., Batruch, I., & Diamandis, E. P. (2009). Mining the ovarian cancer ascites proteome for potential ovarian cancer biomarkers. *Molecular & Cellular Proteomics*, 8(4), 661-669.

Kurt, I., Ture, M., & Kurum, A. T. (2008). Comparing performances of logistic regression, classification and regression tree, and neural networks for predicting coronary artery disease. *Expert Systems with Applications*, 34(1), 366-374. doi: <http://dx.doi.org/10.1016/j.eswa.2006.09.004>.

Li, H., & Sun, J. (2012). Forecasting business failure: The use of nearest-neighbour support vectors and correcting imbalanced samples – Evidence from the Chinese hotel industry. *Tourism Management*, 33(3), 622-634. doi: <http://dx.doi.org/10.1016/j.tourman.2011.07.004>.

Luesley, D., Blackledge, G., Kelly, K., Wade-Evans, T., Fielding, J., Lawton, F., . . . Latief, T. (1988). Failure of second-look laparotomy to influence survival in epithelial ovarian cancer. *The Lancet*, 332(8611), 599-603.

McLean, K. A., Shah, C. A., Thompson, S. A., Gray, H. J., Swensen, R. E., & Goff, B. A. (2010). Ovarian cancer in the elderly: outcomes with neoadjuvant chemotherapy or primary cytoreduction. *Gynecologic Oncology*, 118(1), 43-46.

MedicineNet.com. (2015). Ovarian Cancer Symptoms, Early Warning Signs, and Risk Factors. from <http://www.medicinenet.com/script/main/art.asp?articlekey=47050>.

National Cancer Institute (2015, April 6, 2015). Cancer Statistics. from <http://www.cancer.gov/about-cancer/what-is-cancer/statistics>.

National Ovarian Cancer Coalition. (2015). How is Ovarian Cancer Diagnosed? , from <http://www.ovarian.org/detection.php>.

Society of Gynecologic Oncology. (2014). *International Journal of Gynecology and Obstetrics*.

Sundararajan, V., Hershman, D., Grann, V. R., Jacobson, J. S., & Neugut, A. I. (2002). Variations in the use of chemotherapy for elderly patients with advanced ovarian cancer: a population-based study. *Journal of Clinical Oncology*, 20(1), 173-178.

Ture, M., Kurt, I., Turhan Kurum, A., & Ozdamar, K. (2005). Comparing classification techniques for predicting essential hypertension. *Expert Systems with Applications*, 29(3), 583-588. doi: <http://dx.doi.org/10.1016/j.eswa.2005.04.014>.

U.S. Department of Health & Human Services. Understanding Health Information Privacy. from <http://www.hhs.gov/ocr/privacy/hipaa/understanding/index.html>.

Bellaachia, A., & Guven, E. (2006). Predicting breast cancer survivability using data mining techniques. *Age*, 58(13), 10-110.

Endo, A., Shibata, T., & Tanaka, H. (2008). Comparison of Seven Algorithms to Predict Breast Cancer Survival (< Special Issue> Contribution to 21 Century Intelligent Technologies and Bioinformatics). *Biomedical fuzzy and human sciences: the official journal of the Biomedical Fuzzy Systems Association*, 13(2), 11-16.

Frawley, W., Piatetsky-Shapiro, G., & Matheus, C. (1992). Knowledge Discovery in Databases: An Overview. *AI Magazine*, 13.

Kurt, I., Ture, M., & Kurum, A. T. (2008). Comparing performances of logistic regression, classification and regression tree, and neural networks for predicting coronary artery disease. *Expert Systems with Applications*, 34(1), 366-374. doi: <http://dx.doi.org/10.1016/j.eswa.2006.09.004>.

Ture, M., Kurt, I., Turhan Kurum, A., & Ozdamar, K. (2005). Comparing classification techniques for predicting essential hypertension. *Expert Systems with Applications*, 29(3), 583-588. doi: <http://dx.doi.org/10.1016/j.eswa.2005.04.014>.

Kumar, H., & Bishnoi, A. (2013). Data Mining Techniques: To Anticipate And Adjudicate Breast Cancer Survivability. *International Research journal of Management Science and Technology*, 4(1).

Ahmed, K., Emran, A. A., Jesmin, T., Mukti, R. F., Rahman, M. Z., & Ahmed, F. (2013). Early detection of lung cancer risk using data mining. *Asian Pacific journal of cancer prevention : APJCP*, 14(1), 595-598. doi: 10.7314/APJCP.2013.14.1.595.

Stärk, K. D. C., & Pfeiffer, D. U. (1999). The application of non-parametric techniques to solve classification problems in complex data sets in veterinary epidemiology – An example. *Intelligent Data Analysis*, 3(1), 23-35. doi: [http://dx.doi.org/10.1016/S1088-467X\(99\)00003-7](http://dx.doi.org/10.1016/S1088-467X(99)00003-7).

King, R. D., Feng, C., & Sutherland, A. (1995). Statlog: comparison of classification algorithms on large real-world problems. *Applied Artificial Intelligence an International Journal*, 9(3), 289-333.

## APPENDICES

Name	Variable Number	Type	Format	Label	Length
ADMISSION_TYPE_CODE_DESC	22	Character	\$CHAR		9
Admitted_date	43	Numeric	DATE		8
Bed_size_code	41	Numeric	BEST		8
DAY_NUMBER_OF_WEEK	38	Numeric	BEST		8
DRG_CODE_DESC	39	Character	\$CHAR		45
DRG_TYPE	40	Character	\$CHAR		6
HOLIDAY_IND	37	Numeric	BEST		8
Length_of_stay	45	Numeric	BEST		8
MDC_CODE_DESC	18	Character	\$CHAR		45
MEDICAL_SPECIALTY	20	Character	\$CHAR		25
PATIENT_TYPE_DESC	17	Character	\$CHAR		37
PRESENT_ON_ADMIT_DESC	19	Character	\$CHAR		7
Survival_code	42	Numeric	BEST		8
Unique_patient_id	1	Numeric			8
WEEKDAY_IND	38	Numeric	BEST		8
acute_care	12	Character	\$CHAR		9
admission_source_desc	15	Character	\$CHAR		57
admitted_dt_tm	30	Numeric	DATETIME		8
age_in_years	2	Numeric	BEST		8
bed_size_range	7	Character	\$CHAR		7
caresetting_code_desc	27	Character	\$CHAR		31
cath_lab_bin	9	Numeric	BEST		8
census_division	6	Numeric	BEST		8
census_region	5	Character	\$CHAR		9
diag_cath_lab_bin	10	Numeric	BEST		8
diagnosis_code_desc	14	Character	\$CHAR		61
diagnosis_priority	13	Numeric	BEST		8
discharge_code_desc	16	Character	\$CHAR		18
discharge_date	44	Numeric	DATE		8
discharged_dt_tm	31	Numeric	DATETIME		8
marital_status	4	Character	\$CHAR		17
month	34	Numeric	BEST		8
month_desc	35	Character	\$CHAR		3
payer_code_desc	21	Character	\$CHAR		36
procedure_code	25	Character	\$CHAR		5
procedure_code_desc	26	Character	\$CHAR		255
procedure_dt_tm	29	Numeric	DATETIME		8
procedure_priority	23	Numeric	BEST		8
procedure_type	24	Character	\$CHAR		5
quater	33	Numeric	BEST		8
race	3	Character	\$CHAR		16
teaching_bin	8	Numeric	BEST		8
total_charges	28	Numeric	BEST		8
urban_rural	11	Character	\$CHAR		5
year	32	Numeric	BEST		8

**Table.1** Variable attributes

Comparing \$N-Survival\_code with Survival\_code

'Partition'	1_Training		2_Testing	
Correct	4,410	100%	2,094	96.77%
Wrong	0	0%	70	3.23%
Total	4,410		2,164	

Coincidence Matrix for \$N-Survival\_code (rows show actuals)

'Partition' = 1_Training	0.000000	1.000000	
0.000000	2,135	0	
1.000000	0	2,275	
'Partition' = 2_Testing	0.000000	1.000000	\$null\$
0.000000	1,061	0	0
1.000000	49	1,033	21

Performance Evaluation

'Partition' = 1_Training	
0.000000	0.725
1.000000	0.662
'Partition' = 2_Testing	
0.000000	0.668
1.000000	0.674

**Table.2** Coincidence matrix for balanced data: Neural Network- MLP

Comparing \$N-Survival\_code with Survival\_code

'Partition'	1_Training		2_Testing	
Correct	3,019	68.46%	1,453	67.14%
Wrong	1,391	31.54%	711	32.86%
Total	4,410		2,164	

Coincidence Matrix for \$N-Survival\_code (rows show actuals)

'Partition' = 1_Training	0.000000	1.000000	
0.000000	1,362	773	
1.000000	618	1,657	
'Partition' = 2_Testing	0.000000	1.000000	\$null\$
0.000000	678	383	0
1.000000	307	775	21

Performance Evaluation

'Partition' = 1_Training	
0.000000	0.351
1.000000	0.279
'Partition' = 2_Testing	
0.000000	0.339
1.000000	0.272

**Table.3** Coincidence matrix for balanced data: Neural Network- RBF

Comparing \$C-Survival\_code with Survival\_code

'Partition'	1_Training		2_Testing	
Correct	4,349	98.62%	2,079	96.07%
Wrong	61	1.38%	85	3.93%
Total	4,410		2,164	

Coincidence Matrix for \$C-Survival\_code (rows show actuals)

'Partition' = 1_Training	0.000000	1.000000
0.000000	2,135	0
1.000000	61	2,214
'Partition' = 2_Testing	0.000000	1.000000
0.000000	1,061	0
1.000000	85	1,018

Performance Evaluation

'Partition' = 1_Training	
0.000000	0.697
1.000000	0.662
'Partition' = 2_Testing	
0.000000	0.636
1.000000	0.674

**Table.4** Coincidence matrix for balanced data: C5 Decision tree

Comparing \$N-Survival\_code with Survival\_code

'Partition'	1_Training		2_Testing	
Correct	2,359	100%	1,110	91.96%
Wrong	0	0%	97	8.04%
Total	2,359		1,207	

Coincidence Matrix for \$N-Survival\_code (rows show actuals)

'Partition' = 1_Training	0.000000	1.000000	
0.000000	128	0	
1.000000	0	2,231	
'Partition' = 2_Testing	0.000000	1.000000	\$null\$
0.000000	23	35	2
1.000000	34	1,087	26

Performance Evaluation

'Partition' = 1_Training	
0.000000	2.914
1.000000	0.056
'Partition' = 2_Testing	
0.000000	2.094
1.000000	0.019

**Table.5** Coincidence matrix for unbalanced data: Neural Network- MLP

Comparing \$N-Survival\_code with Survival\_code

'Partition'	1_Training		2_Testing	
Correct	2,231	94.57%	1,121	92.87%
Wrong	128	5.43%	86	7.13%
Total	2,359		1,207	

Coincidence Matrix for \$N-Survival\_code (rows show actuals)

'Partition' = 1_Training	1.000000	
0.000000		128
1.000000		2,231
'Partition' = 2_Testing	1.000000	\$null\$
0.000000		58
1.000000		1,121

Performance Evaluation

'Partition' = 1_Training	
1.000000	0.0
'Partition' = 2_Testing	
1.000000	0.001

**Table.6** Coincidence matrix for unbalanced data: Neural Network- RBF

Comparing \$C-Survival\_code with Survival\_code

'Partition'	1_Training		2_Testing	
Correct	2,246	95.21%	1,145	94.86%
Wrong	113	4.79%	62	5.14%
Total	2,359		1,207	

Coincidence Matrix for \$C-Survival\_code (rows show actuals)

'Partition' = 1_Training	0.000000	1.000000
0.000000		30
1.000000		15
'Partition' = 2_Testing	0.000000	1.000000
0.000000		10
1.000000		12

Performance Evaluation

'Partition' = 1_Training	
0.000000	2.508
1.000000	0.013
'Partition' = 2_Testing	
0.000000	2.213
1.000000	0.008

**Table.7** Coincidence matrix for unbalanced data: C5 Decision tree

Data Role=TRAIN Target=Survival\_code Target Label=' '

False Negative	True Negative	False Positive	True Positive
360	1906	234	1902

Data Role=VALIDATE Target=Survival\_code Target Label=' '

False Negative	True Negative	False Positive	True Positive
172	933	123	944

**Table.8** Coincidence matrix for balanced data: Neural Network- MLP

Data Role=TRAIN Target=Survival\_code Target Label=' '

False Negative	True Negative	False Positive	True Positive
388	2052	88	1874

Data Role=VALIDATE Target=Survival\_code Target Label=' '

False Negative	True Negative	False Positive	True Positive
185	1008	48	931

**Table.9** Coincidence matrix for balanced data: Neural Network- RBF

Data Role=TRAIN Target=Survival\_code Target Label=' '

False Negative	True Negative	False Positive	True Positive
439	2093	47	1823

Data Role=VALIDATE Target=Survival\_code Target Label=' '

False Negative	True Negative	False Positive	True Positive
224	1018	38	892

**Table.10** Coincidence matrix for balanced data: Decision Tree

Data Role=TRAIN Target=Survival_code Target Label=' '			
False Negative	True Negative	False Positive	True Positive
12	41	85	2251

Data Role=VALIDATE Target=Survival_code Target Label=' '			
False Negative	True Negative	False Positive	True Positive
3	20	42	1112

**Table.11** Coincidence matrix for unbalanced data: Neural Network- MLP

Data Role=TRAIN Target=Survival_code Target Label=' '			
False Negative	True Negative	False Positive	True Positive
10	31	95	2253

Data Role=VALIDATE Target=Survival_code Target Label=' '			
False Negative	True Negative	False Positive	True Positive
2	19	43	1113

**Table.12** Coincidence matrix for unbalanced data: Neural Network- RBF

Data Role=TRAIN Target=Survival\_code Target Label=' '

False Negative	True Negative	False Positive	True Positive
6	37	89	2257

Data Role=VALIDATE Target=Survival\_code Target Label=' '

False Negative	True Negative	False Positive	True Positive
3	16	46	1112

**Table.13** Coincidence matrix for unbalanced data: Decision Tree

**Decision tree rules for balanced dataset using neural network MLP**

\*-----\*

Node = 3

\*-----\*

if PATIENT\_TYPE\_DESC IS ONE OF: RECURRING, OUTPATIENT, OUTPATIENT SURGERY, OTHER SPECIALTY, OBSERVATION / SHORT STAY / 24 HR, EMERGENCY, SERIES, DAY SURGERY, RADIOLOGY, CLINIC, OBSERVATION or MISSING

then

Tree Node Identifier = 3

Number of Observations = 1047

Predicted: Survival\_code=1 = 0.98

Predicted: Survival\_code=0 = 0.02

\*-----\*

Node = 9

\*-----\*

if diagnosis\_priority >= 4

AND PATIENT\_TYPE\_DESC IS ONE OF: INPATIENT, PREADMIT

AND ADMISSION\_TYPE\_CODE\_DESC IS ONE OF: ELECTIVE

then

Tree Node Identifier = 9

Number of Observations = 236

Predicted: Survival\_code=1 = 0.24

Predicted: Survival\_code=0 = 0.76

\*-----\*

Node = 10

\*-----\*

if caresetting\_code\_desc IS ONE OF: INTENSIVE CARE UNIT, NOT MAPPED, INTENSIVE CARE UNIT - MEDICAL, CORONARY CARE UNIT, INTENSIVE CARE UNIT - SURGICAL, UROLOGY

AND PATIENT\_TYPE\_DESC IS ONE OF: INPATIENT, PREADMIT

AND ADMISSION\_TYPE\_CODE\_DESC IS ONE OF: URGENT, UNKNOWN, EMERGENCY, NEWBORN or MISSING

then

Tree Node Identifier = 10

Number of Observations = 712

Predicted: Survival\_code=1 = 0.03

Predicted: Survival\_code=0 = 0.97

\*-----\*

Node = 14

\*-----\*

if diagnosis\_priority <= 3 or MISSING

AND caresetting\_code\_desc IS ONE OF: INTENSIVE CARE UNIT, INTENSIVE CARE UNIT - MEDICAL, CORONARY CARE UNIT, INTENSIVE CARE UNIT - SURGICAL, AMBULATORY UNIT, OBSTETRICS & GYNECOLOGY

AND PATIENT\_TYPE\_DESC IS ONE OF: INPATIENT, PREADMIT

AND ADMISSION\_TYPE\_CODE\_DESC IS ONE OF: ELECTIVE

then

Tree Node Identifier = 14

Number of Observations = 87

Predicted: Survival\_code=1 = 0.13

Predicted: Survival\_code=0 = 0.87

\*-----\*

Node = 15

\*-----\*

if diagnosis\_priority <= 3 or MISSING

AND caresetting\_code\_desc IS ONE OF: CARE SETTING UNDEFINED, ONCOLOGY,  
GYNECOLOGY, NOT MAPPED, OBSTETRICS, MEDICAL/SURGICAL, SURGERY,  
AMBULATORY SURGERY, NURSING HOME (LTC), ONCOLOGY - GYNECOLOGY,  
ORT

AND PATIENT\_TYPE\_DESC IS ONE OF: INPATIENT, PREADMIT

AND ADMISSION\_TYPE\_CODE\_DESC IS ONE OF: ELECTIVE

then

Tree Node Identifier = 15

Number of Observations = 791

Predicted: Survival\_code=1 = 0.86

Predicted: Survival\_code=0 = 0.14

\*-----\*

Node = 21

\*-----\*

if caresetting\_code\_desc IS ONE OF: CARE SETTING UNDEFINED, ONCOLOGY,  
MEDICAL/SURGICAL, MULTISPECIALTY UNIT, SURGERY, CARDIOLOGY,  
EMERGENCY ROOM, ONCOLOGY - GYNECOLOGY, ORTHOPEDICS, PEDIATRICS,  
NOT A,

AND PATIENT\_TYPE\_DESC IS ONE OF: INPATIENT, PREADMIT

AND ADMISSION\_TYPE\_CODE\_DESC IS ONE OF: EMERGENCY, NEWBORN or  
MISSING

then

Tree Node Identifier = 21

Number of Observations = 884

Predicted: Survival\_code=1 = 0.21

Predicted: Survival\_code=0 = 0.79

\*-----\*

Node = 32

\*-----\*

if payer\_code\_desc IS ONE OF: MEDICARE, HMO/MANAGED CARE (UNDESIGNATED),  
BLUE CROSS/BLUE SHIELD, MEDICAID, MEDICAID MANAGED CARE (UNDESIGNA,  
MEDICARE MANAGED CARE (UNDESIGNA, OTHER GOVERNMENT

AND caresetting\_code\_desc IS ONE OF: CARE SETTING UNDEFINED, ONCOLOGY,  
MEDICAL/SURGICAL, MULTISPECIALTY UNIT, SURGERY, CARDIOLOGY,  
EMERGENCY ROOM, ONCOLOGY - GYNECOLOGY, ORTHOPEDICS, PEDIATRICS,  
NOT A

AND PATIENT\_TYPE\_DESC IS ONE OF: INPATIENT, PREADMIT

AND ADMISSION\_TYPE\_CODE\_DESC IS ONE OF: URGENT, UNKNOWN

then

Tree Node Identifier = 32

Number of Observations = 539

Predicted: Survival\_code=1 = 0.36

Predicted: Survival\_code=0 = 0.64

\*-----\*

Node = 33

\*-----\*

if payer\_code\_desc IS ONE OF: PPO (UNDESIGNATED), OTHER COMMERCIAL PAYER,  
OTHER NON-GOVT, SELF-PAY or MISSING

AND caresetting\_code\_desc IS ONE OF: CARE SETTING UNDEFINED, ONCOLOGY,  
MEDICAL/SURGICAL, MULTISPECIALTY UNIT, SURGERY, CARDIOLOGY,

EMERGENCY ROOM, ONCOLOGY - GYNECOLOGY, ORTHOPEDICS, PEDIATRICS,  
NOT A

AND PATIENT\_TYPE\_DESC IS ONE OF: INPATIENT, PREADMIT

AND ADMISSION\_TYPE\_CODE\_DESC IS ONE OF: URGENT, UNKNOWN

then

Tree Node Identifier = 33

Number of Observations = 106

Predicted: Survival\_code=1 = 0.83

Predicted: Survival\_code=0 = 0.17

**Decision tree rules for balanced dataset using neural network RBF**

\*-----\*

Node = 3

\*-----\*

if PATIENT\_TYPE\_DESC IS ONE OF: RECURRING, OUTPATIENT, OUTPATIENT  
SURGERY, OTHER SPECIALTY, OBSERVATION / SHORT STAY / 24 HR, EMERGENCY,  
SERIES, DAY SURGERY, RADIOLOGY, CLINIC, OBSERVATION or MISSING

then

Tree Node Identifier = 3

Number of Observations = 1047

Predicted: Survival\_code=1 = 0.98

Predicted: Survival\_code=0 = 0.02

\*-----\*

Node = 9

\*-----\*

if diagnosis\_priority >= 4

AND PATIENT\_TYPE\_DESC IS ONE OF: INPATIENT, PREADMIT

AND ADMISSION\_TYPE\_CODE\_DESC IS ONE OF: ELECTIVE

then

Tree Node Identifier = 9

Number of Observations = 236

Predicted: Survival\_code=1 = 0.24

Predicted: Survival\_code=0 = 0.76

\*-----\*

Node = 10

\*-----\*

if caresetting\_code\_desc IS ONE OF: INTENSIVE CARE UNIT, NOT MAPPED, INTENSIVE CARE UNIT - MEDICAL, CORONARY CARE UNIT, INTENSIVE CARE UNIT - SURGICAL, UROLOGY

AND PATIENT\_TYPE\_DESC IS ONE OF: INPATIENT, PREADMIT

AND ADMISSION\_TYPE\_CODE\_DESC IS ONE OF: URGENT, UNKNOWN, EMERGENCY, NEWBORN or MISSING

then

Tree Node Identifier = 10

Number of Observations = 712

Predicted: Survival\_code=1 = 0.03

Predicted: Survival\_code=0 = 0.97

\*-----\*

Node = 14

\*-----\*

if diagnosis\_priority <= 3 or MISSING

AND caresetting\_code\_desc IS ONE OF: INTENSIVE CARE UNIT, INTENSIVE CARE UNIT - MEDICAL, CORONARY CARE UNIT, INTENSIVE CARE UNIT - SURGICAL, AMBULATORY UNIT, OBSTETRICS & GYNECOLOGY

AND PATIENT\_TYPE\_DESC IS ONE OF: INPATIENT, PREADMIT

AND ADMISSION\_TYPE\_CODE\_DESC IS ONE OF: ELECTIVE

then

Tree Node Identifier = 14

Number of Observations = 87

Predicted: Survival\_code=1 = 0.13

Predicted: Survival\_code=0 = 0.87

\*-----\*

Node = 15

\*-----\*

if diagnosis\_priority <= 3 or MISSING

AND caresetting\_code\_desc IS ONE OF: CARE SETTING UNDEFINED, ONCOLOGY,  
GYNECOLOGY, NOT MAPPED, OBSTETRICS, MEDICAL/SURGICAL, SURGERY,  
AMBULATORY SURGERY, NURSING HOME (LTC), ONCOLOGY - GYNECOLOGY,  
ORT

AND PATIENT\_TYPE\_DESC IS ONE OF: INPATIENT, PREADMIT

AND ADMISSION\_TYPE\_CODE\_DESC IS ONE OF: ELECTIVE

then

Tree Node Identifier = 15

Number of Observations = 791

Predicted: Survival\_code=1 = 0.86

Predicted: Survival\_code=0 = 0.14

\*-----\*

Node = 21

\*-----\*

if caresetting\_code\_desc IS ONE OF: CARE SETTING UNDEFINED, ONCOLOGY, MEDICAL/SURGICAL, MULTISPECIALTY UNIT, SURGERY, CARDIOLOGY, EMERGENCY ROOM, ONCOLOGY - GYNECOLOGY, ORTHOPEDICS, PEDIATRICS, NOT A,

AND PATIENT\_TYPE\_DESC IS ONE OF: INPATIENT, PREADMIT

AND ADMISSION\_TYPE\_CODE\_DESC IS ONE OF: EMERGENCY, NEWBORN or MISSING

then

Tree Node Identifier = 21

Number of Observations = 884

Predicted: Survival\_code=1 = 0.21

Predicted: Survival\_code=0 = 0.79

\*-----\*

Node = 32

\*-----\*

if payer\_code\_desc IS ONE OF: MEDICARE, HMO/MANAGED CARE (UNDESIGNATED), BLUE CROSS/BLUE SHIELD, MEDICAID, MEDICAID MANAGED CARE (UNDESIGNA, MEDICARE MANAGED CARE (UNDESIGNA, OTHER GOVERNMENT

AND caresetting\_code\_desc IS ONE OF: CARE SETTING UNDEFINED, ONCOLOGY, MEDICAL/SURGICAL, MULTISPECIALTY UNIT, SURGERY, CARDIOLOGY, EMERGENCY ROOM, ONCOLOGY - GYNECOLOGY, ORTHOPEDICS, PEDIATRICS, NOT A

AND PATIENT\_TYPE\_DESC IS ONE OF: INPATIENT, PREADMIT

AND ADMISSION\_TYPE\_CODE\_DESC IS ONE OF: URGENT, UNKNOWN

then

Tree Node Identifier = 32

Number of Observations = 539

Predicted: Survival\_code=1 = 0.36

Predicted: Survival\_code=0 = 0.64

\*-----\*

Node = 33

\*-----\*

if payer\_code\_desc IS ONE OF: PPO (UNDESIGNATED), OTHER COMMERCIAL PAYER,  
OTHER NON-GOVT, SELF-PAY or MISSING

AND caresetting\_code\_desc IS ONE OF: CARE SETTING UNDEFINED, ONCOLOGY,  
MEDICAL/SURGICAL, MULTISPECIALTY UNIT, SURGERY, CARDIOLOGY,  
EMERGENCY ROOM, ONCOLOGY - GYNECOLOGY, ORTHOPEDICS, PEDIATRICS,  
NOT A

AND PATIENT\_TYPE\_DESC IS ONE OF: INPATIENT, PREADMIT

AND ADMISSION\_TYPE\_CODE\_DESC IS ONE OF: URGENT, UNKNOWN

then

Tree Node Identifier = 33

Number of Observations = 106

Predicted: Survival\_code=1 = 0.83

Predicted: Survival\_code=0 = 0.17

**Decision tree rules for balanced dataset using Decision tree**

\*-----\*

Node = 7

\*-----\*

if admission\_source\_desc IS ONE OF: PHYSICIAN REFERRAL, CLINIC REFERRAL,  
UNKNOWN or MISSING

AND PATIENT\_TYPE\_DESC IS ONE OF: RECURRING, OUTPATIENT, OUTPATIENT  
SURGERY, OTHER SPECIALTY, OBSERVATION / SHORT STAY / 24 HR, EMERGENCY,  
SERIES, DAY SURGERY, RADIOLOGY, CLINIC, OBSERVATION or MISSING

then

Tree Node Identifier = 7

Number of Observations = 1003

Predicted: Survival\_code=1 = 1.00

Predicted: Survival\_code=0 = 0.00

\*-----\*

Node = 10

\*-----\*

if caresetting\_code\_desc IS ONE OF: INTENSIVE CARE UNIT, NOT MAPPED, INTENSIVE CARE UNIT - MEDICAL, CORONARY CARE UNIT, INTENSIVE CARE UNIT - SURGICAL, UROLOGY

AND PATIENT\_TYPE\_DESC IS ONE OF: INPATIENT, PREADMIT

AND ADMISSION\_TYPE\_CODE\_DESC IS ONE OF: URGENT, UNKNOWN, EMERGENCY, NEWBORN or MISSING

then

Tree Node Identifier = 10

Number of Observations = 712

Predicted: Survival\_code=1 = 0.03

Predicted: Survival\_code=0 = 0.97

\*-----\*

Node = 12

\*-----\*

if census\_division IS ONE OF: 5

AND admission\_source\_desc IS ONE OF: EMERGENCY ROOM, TRANSFER FROM A SKILLED NURSING

AND PATIENT\_TYPE\_DESC IS ONE OF: RECURRING, OUTPATIENT, OUTPATIENT SURGERY, OTHER SPECIALTY, OBSERVATION / SHORT STAY / 24 HR, EMERGENCY, SERIES, DAY SURGERY, RADIOLOGY, CLINIC, OBSERVATION or MISSING

then

Tree Node Identifier = 12

Number of Observations = 23

Predicted: Survival\_code=1 = 0.13

Predicted: Survival\_code=0 = 0.87

\*-----\*

Node = 13

\*-----\*

if census\_division IS ONE OF: 4, 1 or MISSING

AND admission\_source\_desc IS ONE OF: EMERGENCY ROOM, TRANSFER FROM A SKILLED NURSING

AND PATIENT\_TYPE\_DESC IS ONE OF: RECURRING, OUTPATIENT, OUTPATIENT SURGERY, OTHER SPECIALTY, OBSERVATION / SHORT STAY / 24 HR, EMERGENCY, SERIES, DAY SURGERY, RADIOLOGY, CLINIC, OBSERVATION or MISSING

then

Tree Node Identifier = 13

Number of Observations = 21

Predicted: Survival\_code=1 = 1.00

Predicted: Survival\_code=0 = 0.00

\*-----\*

Node = 16

\*-----\*

if diagnosis\_priority >= 4

AND age\_in\_years < 55.5

AND PATIENT\_TYPE\_DESC IS ONE OF: INPATIENT, PREADMIT

AND ADMISSION\_TYPE\_CODE\_DESC IS ONE OF: ELECTIVE

then

Tree Node Identifier = 16

Number of Observations = 23

Predicted: Survival\_code=1 = 1.00

Predicted: Survival\_code=0 = 0.00

\*-----\*

Node = 22

\*-----\*

if year IS ONE OF: 2011, 2010, 2008, 2005, 2006, 2013

AND diagnosis\_priority <= 3 or MISSING

AND caresetting\_code\_desc IS ONE OF: INTENSIVE CARE UNIT, INTENSIVE CARE UNIT  
- MEDICAL, CORONARY CARE UNIT, INTENSIVE CARE UNIT - SURGICAL,  
AMBULATORY UNIT, OBSTETRICS & GYNECOLOGY

AND PATIENT\_TYPE\_DESC IS ONE OF: INPATIENT, PREADMIT

AND ADMISSION\_TYPE\_CODE\_DESC IS ONE OF: ELECTIVE

then

Tree Node Identifier = 22

Number of Observations = 81

Predicted: Survival\_code=1 = 0.06

Predicted: Survival\_code=0 = 0.94

\*-----\*

Node = 23

\*-----\*

if year equals Missing

AND diagnosis\_priority <= 3 or MISSING

AND caresetting\_code\_desc IS ONE OF: INTENSIVE CARE UNIT, INTENSIVE CARE UNIT - MEDICAL, CORONARY CARE UNIT, INTENSIVE CARE UNIT - SURGICAL, AMBULATORY UNIT, OBSTETRICS & GYNECOLOGY

AND PATIENT\_TYPE\_DESC IS ONE OF: INPATIENT, PREADMIT

AND ADMISSION\_TYPE\_CODE\_DESC IS ONE OF: ELECTIVE

then

Tree Node Identifier = 23

Number of Observations = 6

Predicted: Survival\_code=1 = 1.00

Predicted: Survival\_code=0 = 0.00

\*-----\*

Node = 26

\*-----\*

if diagnosis\_priority >= 4

AND caresetting\_code\_desc IS ONE OF: CARE SETTING UNDEFINED, ONCOLOGY, INTENSIVE CARE UNIT, NOT MAPPED, NURSING HOME (LTC), ORTHOPEDICS, INTENSIVE CARE UNIT - NEUROLOGY

AND age\_in\_years >= 55.5 or MISSING

AND PATIENT\_TYPE\_DESC IS ONE OF: INPATIENT, PREADMIT

AND ADMISSION\_TYPE\_CODE\_DESC IS ONE OF: ELECTIVE

then

Tree Node Identifier = 26

Number of Observations = 203

Predicted: Survival\_code=1 = 0.11

Predicted: Survival\_code=0 = 0.89

\*-----\*

Node = 27

\*-----\*

if diagnosis\_priority >= 4

AND caresetting\_code\_desc equals Missing

AND age\_in\_years >= 55.5 or MISSING

AND PATIENT\_TYPE\_DESC IS ONE OF: INPATIENT, PREADMIT

AND ADMISSION\_TYPE\_CODE\_DESC IS ONE OF: ELECTIVE

then

Tree Node Identifier = 27

Number of Observations = 10

Predicted: Survival\_code=1 = 1.00

Predicted: Survival\_code=0 = 0.00

\*-----\*

Node = 32

\*-----\*

if payer\_code\_desc IS ONE OF: MEDICARE, HMO/MANAGED CARE (UNDESIGNATED),  
BLUE CROSS/BLUE SHIELD, MEDICAID, MEDICAID MANAGED CARE (UNDESIGNA,  
MEDICARE MANAGED CARE (UNDESIGNA, OTHER GOVERNMENT

AND caresetting\_code\_desc IS ONE OF: CARE SETTING UNDEFINED, ONCOLOGY,  
MEDICAL/SURGICAL, MULTISPECIALTY UNIT, SURGERY, CARDIOLOGY,  
EMERGENCY ROOM, ONCOLOGY - GYNECOLOGY, ORTHOPEDICS, PEDIATRICS,  
NOT A

AND PATIENT\_TYPE\_DESC IS ONE OF: INPATIENT, PREADMIT

AND ADMISSION\_TYPE\_CODE\_DESC IS ONE OF: URGENT, UNKNOWN

then

Tree Node Identifier = 32

Number of Observations = 539

Predicted: Survival\_code=1 = 0.36

Predicted: Survival\_code=0 = 0.64

\*-----\*

Node = 33

\*-----\*

if payer\_code\_desc IS ONE OF: PPO (UNDESIGNATED), OTHER COMMERCIAL PAYER,  
OTHER NON-GOVT, SELF-PAY or MISSING

AND caresetting\_code\_desc IS ONE OF: CARE SETTING UNDEFINED, ONCOLOGY,  
MEDICAL/SURGICAL, MULTISPECIALTY UNIT, SURGERY, CARDIOLOGY,  
EMERGENCY ROOM, ONCOLOGY - GYNECOLOGY, ORTHOPEDICS, PEDIATRICS,  
NOT A

AND PATIENT\_TYPE\_DESC IS ONE OF: INPATIENT, PREADMIT

AND ADMISSION\_TYPE\_CODE\_DESC IS ONE OF: URGENT, UNKNOWN

then

Tree Node Identifier = 33

Number of Observations = 106

Predicted: Survival\_code=1 = 0.83

Predicted: Survival\_code=0 = 0.17

\*-----\*

Node = 38

\*-----\*

if diagnosis\_priority <= 3 or MISSING

AND caresetting\_code\_desc IS ONE OF: CARE SETTING UNDEFINED, ONCOLOGY,  
GYNECOLOGY, NOT MAPPED, OBSTETRICS, MEDICAL/SURGICAL, SURGERY,  
AMBULATORY SURGERY, NURSING HOME (LTC), ONCOLOGY - GYNECOLOGY,  
ORT

AND age\_in\_years < 66.5 or MISSING

AND PATIENT\_TYPE\_DESC IS ONE OF: INPATIENT, PREADMIT

AND Length\_of\_stay < 10.5 or MISSING

AND ADMISSION\_TYPE\_CODE\_DESC IS ONE OF: ELECTIVE

then

Tree Node Identifier = 38

Number of Observations = 475

Predicted: Survival\_code=1 = 1.00

Predicted: Survival\_code=0 = 0.00

\*-----\*

Node = 39

\*-----\*

if diagnosis\_priority <= 3 or MISSING

AND caresetting\_code\_desc IS ONE OF: CARE SETTING UNDEFINED, ONCOLOGY,  
GYNECOLOGY, NOT MAPPED, OBSTETRICS, MEDICAL/SURGICAL, SURGERY,  
AMBULATORY SURGERY, NURSING HOME (LTC), ONCOLOGY - GYNECOLOGY,  
ORT

AND age\_in\_years < 66.5 or MISSING

AND PATIENT\_TYPE\_DESC IS ONE OF: INPATIENT, PREADMIT

AND Length\_of\_stay >= 10.5

AND ADMISSION\_TYPE\_CODE\_DESC IS ONE OF: ELECTIVE

then

Tree Node Identifier = 39

Number of Observations = 29

Predicted: Survival\_code=1 = 0.31

Predicted: Survival\_code=0 = 0.69

\*-----\*

Node = 40

\*-----\*

if diagnosis\_priority <= 3 or MISSING

AND census\_division IS ONE OF: 2, 3, 5 or MISSING

AND caresetting\_code\_desc IS ONE OF: CARE SETTING UNDEFINED, ONCOLOGY, GYNECOLOGY, NOT MAPPED, OBSTETRICS, MEDICAL/SURGICAL, SURGERY, AMBULATORY SURGERY, NURSING HOME (LTC), ONCOLOGY - GYNECOLOGY, ORT

AND age\_in\_years >= 66.5

AND PATIENT\_TYPE\_DESC IS ONE OF: INPATIENT, PREADMIT

AND ADMISSION\_TYPE\_CODE\_DESC IS ONE OF: ELECTIVE

then

Tree Node Identifier = 40

Number of Observations = 163

Predicted: Survival\_code=1 = 0.44

Predicted: Survival\_code=0 = 0.56

\*-----\*

Node = 41

\*-----\*

if diagnosis\_priority <= 3 or MISSING

AND census\_division IS ONE OF: 6, 4, 7, 8, 1, 9

AND caresetting\_code\_desc IS ONE OF: CARE SETTING UNDEFINED, ONCOLOGY, GYNECOLOGY, NOT MAPPED, OBSTETRICS, MEDICAL/SURGICAL, SURGERY, AMBULATORY SURGERY, NURSING HOME (LTC), ONCOLOGY - GYNECOLOGY, ORT

AND age\_in\_years >= 66.5

AND PATIENT\_TYPE\_DESC IS ONE OF: INPATIENT, PREADMIT

AND ADMISSION\_TYPE\_CODE\_DESC IS ONE OF: ELECTIVE

then

Tree Node Identifier = 41

Number of Observations = 124

Predicted: Survival\_code=1 = 1.00

Predicted: Survival\_code=0 = 0.00

\*-----\*

Node = 52

\*-----\*

if diagnosis\_code\_desc IS ONE OF: MALIGNANT NEOPLASM OF OVARY, MALIGNANT  
NEOPLASM OF PARAMETRIU

AND caresetting\_code\_desc IS ONE OF: CARE SETTING UNDEFINED, CARDIOLOGY,  
EMERGENCY ROOM, NOT A CARE SETTING, PULMONOLOGY

AND PATIENT\_TYPE\_DESC IS ONE OF: INPATIENT, PREADMIT

AND ADMISSION\_TYPE\_CODE\_DESC IS ONE OF: EMERGENCY, NEWBORN or  
MISSING

then

Tree Node Identifier = 52

Number of Observations = 726

Predicted: Survival\_code=1 = 0.14

Predicted: Survival\_code=0 = 0.86

\*-----\*

Node = 53

\*-----\*

if diagnosis\_code\_desc equals Missing

AND caresetting\_code\_desc IS ONE OF: CARE SETTING UNDEFINED, CARDIOLOGY,  
EMERGENCY ROOM, NOT A CARE SETTING, PULMONOLOGY

AND PATIENT\_TYPE\_DESC IS ONE OF: INPATIENT, PREADMIT

AND ADMISSION\_TYPE\_CODE\_DESC IS ONE OF: EMERGENCY, NEWBORN or MISSING

then

Tree Node Identifier = 53

Number of Observations = 7

Predicted: Survival\_code=1 = 1.00

Predicted: Survival\_code=0 = 0.00

\*-----\*

Node = 54

\*-----\*

if caresetting\_code\_desc IS ONE OF: ONCOLOGY, MEDICAL/SURGICAL,  
MULTISPECIALTY UNIT, SURGERY, ORTHOPEDICS, STEP-DOWN UNIT or MISSING

AND age\_in\_years < 79.5 or MISSING

AND PATIENT\_TYPE\_DESC IS ONE OF: INPATIENT, PREADMIT

AND ADMISSION\_TYPE\_CODE\_DESC IS ONE OF: EMERGENCY, NEWBORN or MISSING

then

Tree Node Identifier = 54

Number of Observations = 95

Predicted: Survival\_code=1 = 0.69

Predicted: Survival\_code=0 = 0.31

\*-----\*

Node = 55

\*-----\*

if caresetting\_code\_desc IS ONE OF: ONCOLOGY, MEDICAL/SURGICAL,  
MULTISPECIALTY UNIT, SURGERY, ORTHOPEDICS, STEP-DOWN UNIT or MISSING

AND age\_in\_years >= 79.5

AND PATIENT\_TYPE\_DESC IS ONE OF: INPATIENT, PREADMIT

AND ADMISSION\_TYPE\_CODE\_DESC IS ONE OF: EMERGENCY, NEWBORN or  
MISSING

then

Tree Node Identifier = 55

Number of Observations = 56

Predicted: Survival\_code=1 = 0.07

Predicted: Survival\_code=0 = 0.93

### **Decision tree rules for unbalanced dataset using neural network MLP**

-----\*

Node = 2

\*-----\*

if caresetting\_code\_desc IS ONE OF: INTENSIVE CARE UNIT, INTENSIVE CARE UNIT -  
MEDICAL, INTENSIVE CARE UNIT - SURGICAL, CORONARY CARE UNIT

then

Tree Node Identifier = 2

Number of Observations = 43

Predicted: Survival\_code=1 = 0.14

Predicted: Survival\_code=0 = 0.86

\*-----\*

Node = 7

\*-----\*

if caresetting\_code\_desc IS ONE OF: CARE SETTING UNDEFINED, ONCOLOGY,  
GYNECOLOGY, MULTISPECIALTY UNIT, NOT MAPPED, OBSTETRICS,  
MEDICAL/SURGICAL, SURGERY, OUTPATIENT CLINIC, AMBULATORY SURGERY,  
AMBULAT

AND Length\_of\_stay < 9.5 AND Length\_of\_stay >= 5.5

then

Tree Node Identifier = 7

Number of Observations = 299

Predicted: Survival\_code=1 = 0.93

Predicted: Survival\_code=0 = 0.07

\*-----\*

Node = 8

\*-----\*

if caresetting\_code\_desc IS ONE OF: NOT MAPPED

AND Length\_of\_stay >= 9.5

then

Tree Node Identifier = 8

Number of Observations = 9

Predicted: Survival\_code=1 = 0.22

Predicted: Survival\_code=0 = 0.78

\*-----\*

Node = 9

\*-----\*

if caresetting\_code\_desc IS ONE OF: CARE SETTING UNDEFINED, ONCOLOGY,  
MEDICAL/SURGICAL, SURGERY, ONCOLOGY - GYNECOLOGY or MISSING

AND Length\_of\_stay >= 9.5

then

Tree Node Identifier = 9

Number of Observations = 149

Predicted: Survival\_code=1 = 0.82

Predicted: Survival\_code=0 = 0.18

\*-----\*

Node = 10

\*-----\*

if caresetting\_code\_desc IS ONE OF: CARE SETTING UNDEFINED, ONCOLOGY,  
GYNECOLOGY, MULTISPECIALTY UNIT, NOT MAPPED, OBSTETRICS,  
MEDICAL/SURGICAL, SURGERY, OUTPATIENT CLINIC, AMBULATORY SURGERY,  
AMBULAT

AND Length\_of\_stay < 0.5 or MISSING

then

Tree Node Identifier = 10

Number of Observations = 964

Predicted: Survival\_code=1 = 1.00

Predicted: Survival\_code=0 = 0.00

\*-----\*

Node = 11

\*-----\*

if caresetting\_code\_desc IS ONE OF: CARE SETTING UNDEFINED, ONCOLOGY,  
GYNECOLOGY, MULTISPECIALTY UNIT, NOT MAPPED, OBSTETRICS,  
MEDICAL/SURGICAL, SURGERY, OUTPATIENT CLINIC, AMBULATORY SURGERY,  
AMBULAT

AND Length\_of\_stay < 5.5 AND Length\_of\_stay >= 0.5

then

Tree Node Identifier = 11

Number of Observations = 925

Predicted: Survival\_code=1 = 0.97

Predicted: Survival\_code=0 = 0.03

**Decision tree rules for unbalanced dataset using neural network RBF**

\*-----\*

Node = 2

\*-----\*

if caresetting\_code\_desc IS ONE OF: INTENSIVE CARE UNIT, INTENSIVE CARE UNIT -  
MEDICAL, INTENSIVE CARE UNIT - SURGICAL, CORONARY CARE UNIT

then

Tree Node Identifier = 2

Number of Observations = 43

Predicted: Survival\_code=1 = 0.14

Predicted: Survival\_code=0 = 0.86

\*-----\*

Node = 7

\*-----\*

if caresetting\_code\_desc IS ONE OF: CARE SETTING UNDEFINED, ONCOLOGY,  
GYNECOLOGY, MULTISPECIALTY UNIT, NOT MAPPED, OBSTETRICS,  
MEDICAL/SURGICAL, SURGERY, OUTPATIENT CLINIC, AMBULATORY SURGERY,  
AMBULAT

AND Length\_of\_stay < 9.5 AND Length\_of\_stay >= 5.5

then

Tree Node Identifier = 7

Number of Observations = 299

Predicted: Survival\_code=1 = 0.93

Predicted: Survival\_code=0 = 0.07

\*-----\*

Node = 8

\*-----\*

if caresetting\_code\_desc IS ONE OF: NOT MAPPED

AND Length\_of\_stay >= 9.5

then

Tree Node Identifier = 8

Number of Observations = 9

Predicted: Survival\_code=1 = 0.22

Predicted: Survival\_code=0 = 0.78

\*-----\*

Node = 9

\*-----\*

if caresetting\_code\_desc IS ONE OF: CARE SETTING UNDEFINED, ONCOLOGY,  
MEDICAL/SURGICAL, SURGERY, ONCOLOGY - GYNECOLOGY or MISSING

AND Length\_of\_stay >= 9.5

then

Tree Node Identifier = 9

Number of Observations = 149

Predicted: Survival\_code=1 = 0.82

Predicted: Survival\_code=0 = 0.18

\*-----\*

Node = 10

\*-----\*

if caresetting\_code\_desc IS ONE OF: CARE SETTING UNDEFINED, ONCOLOGY,  
GYNECOLOGY, MULTISPECIALTY UNIT, NOT MAPPED, OBSTETRICS,  
MEDICAL/SURGICAL, SURGERY, OUTPATIENT CLINIC, AMBULATORY SURGERY,  
AMBULAT

AND Length\_of\_stay < 0.5 or MISSING

then

Tree Node Identifier = 10

Number of Observations = 964

Predicted: Survival\_code=1 = 1.00

Predicted: Survival\_code=0 = 0.00

\*-----\*

Node = 11

\*-----\*

if caresetting\_code\_desc IS ONE OF: CARE SETTING UNDEFINED, ONCOLOGY,  
GYNECOLOGY, MULTISPECIALTY UNIT, NOT MAPPED, OBSTETRICS,  
MEDICAL/SURGICAL, SURGERY, OUTPATIENT CLINIC, AMBULATORY SURGERY,  
AMBULAT

AND Length\_of\_stay < 5.5 AND Length\_of\_stay >= 0.5

then

Tree Node Identifier = 11

Number of Observations = 925

Predicted: Survival\_code=1 = 0.97

Predicted: Survival\_code=0 = 0.03

**Decision tree rules for unbalanced dataset using decision tree**

Node = 2

\*-----\*

if caresetting\_code\_desc IS ONE OF: INTENSIVE CARE UNIT, INTENSIVE CARE UNIT - MEDICAL, INTENSIVE CARE UNIT - SURGICAL, CORONARY CARE UNIT

then

Tree Node Identifier = 2

Number of Observations = 43

Predicted: Survival\_code=1 = 0.14

Predicted: Survival\_code=0 = 0.86

\*-----\*

Node = 3

\*-----\*

if caresetting\_code\_desc IS ONE OF: CARE SETTING UNDEFINED, ONCOLOGY, GYNECOLOGY, MULTISPECIALTY UNIT, NOT MAPPED, OBSTETRICS, MEDICAL/SURGICAL, SURGERY, OUTPATIENT CLINIC, AMBULATORY SURGERY, AMBULAT

then

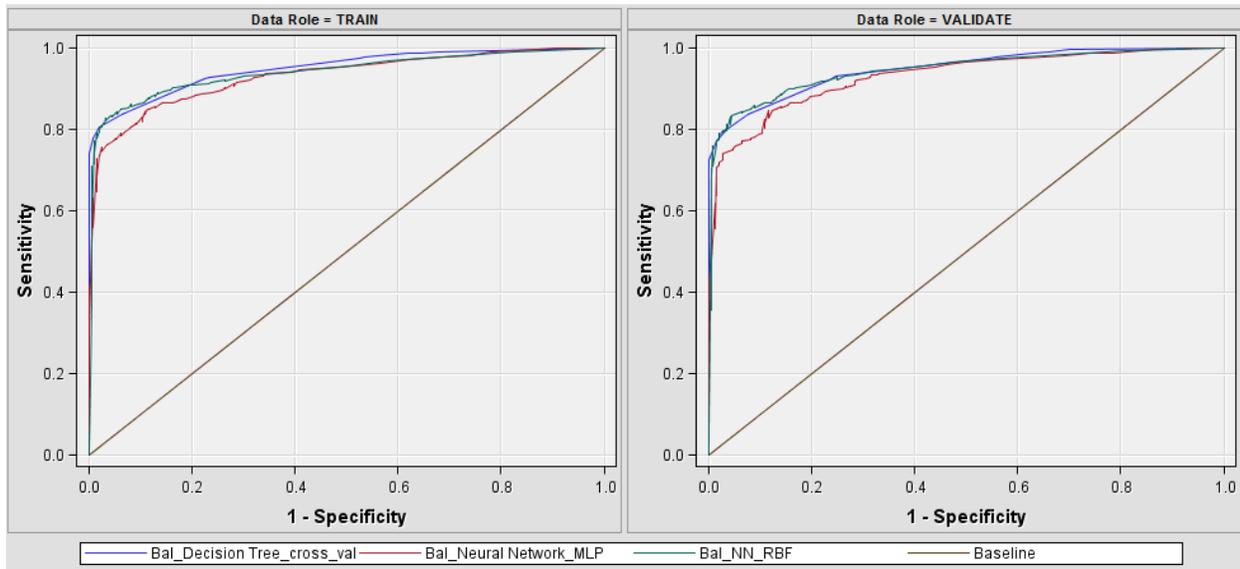
Tree Node Identifier = 3

Number of Observations = 2346

Predicted: Survival\_code=1 = 0.96

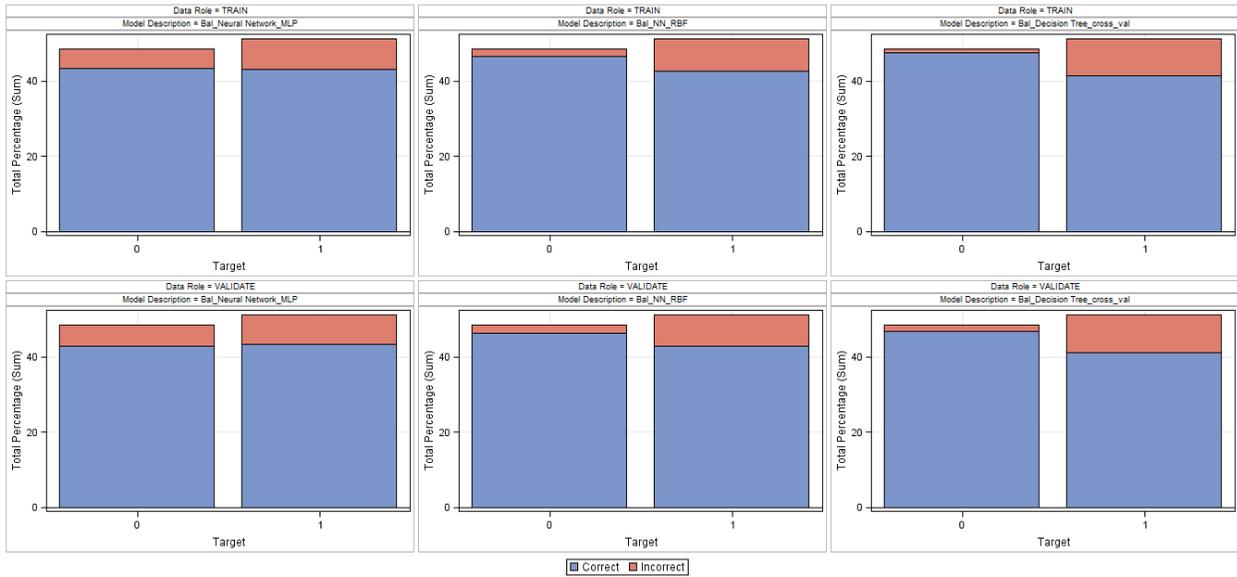
Predicted: Survival\_code=0 = 0.04

## Model Comparison: Balanced Data

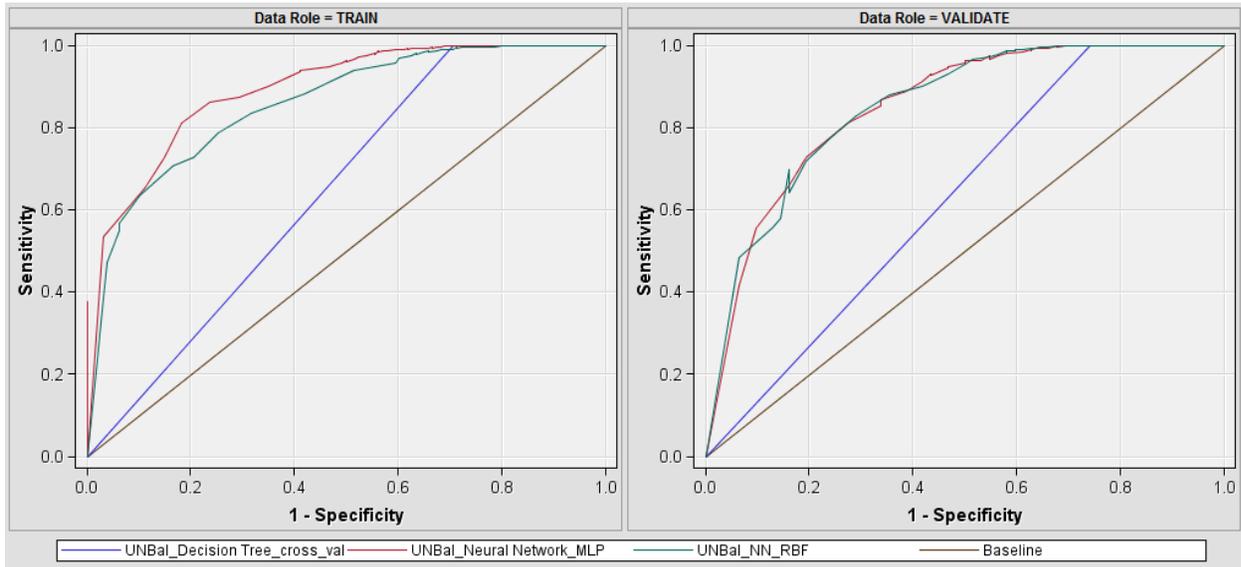


Selected Model	Model Node	Model Description	Valid: Misclassification Rate
Y	Neural2	Bal_NN_RBF	0.10727
	Tree	Bal_Decision Tree_cross_val	0.12063
	Neural5	Bal_Neural Network_MLP	0.13582

Model Description	Data Role	False Negative	True Negative	False Positive	True Positive
Bal_Neural Network_MLP	TRAIN	360	1906	234	1902
Bal_Neural Network_MLP	VALIDATE	172	933	123	944
Bal_NN_RBF	TRAIN	388	2052	88	1874
Bal_NN_RBF	VALIDATE	185	1008	48	931
Bal_Decision Tree_cross_val	TRAIN	439	2093	47	1823
Bal_Decision Tree_cross_val	VALIDATE	224	1018	38	892

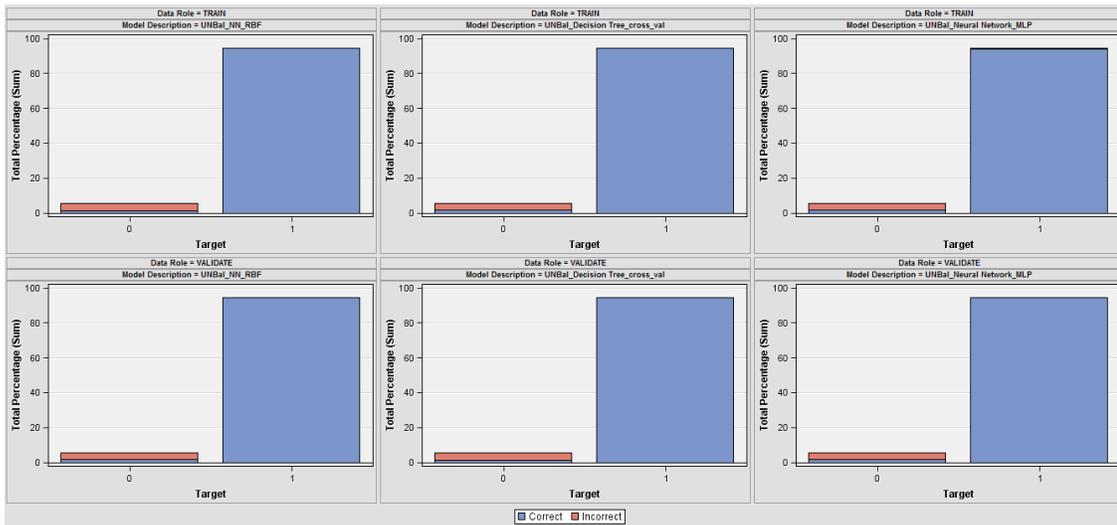


### Model Comparison: UnBalanced Data



Selected Model	Model Node	Model Description	Valid: Misclassification Rate
Y	Neural	UNBal_NN_RBF	0.038233
	Neural4	UNBal_Neural Network_MLP	0.038233
	Tree3	UNBal_Decision Tree_cross_val	0.041631

Model Description	Data Role	False Negative	True Negative	False Positive	True Positive
UNBal_MN_RBF	TRAIN	10	31	95	2253
UNBal_MN_RBF	VALIDATE	2	19	43	1113
UNBal_Decision Tree_cross_val	TRAIN	6	37	89	2257
UNBal_Decision Tree_cross_val	VALIDATE	3	16	46	1112
UNBal_Neural Network_MLP	TRAIN	12	41	85	2251
UNBal_Neural Network_MLP	VALIDATE	3	20	42	1112



VITA

VEDIKA HARISHKUMAR DENGADA

Candidate for the Degree of

Master of Science

Thesis: APPLICATION OF MACHINE LEARNING IN PREDICTING OVARIAN  
CANCER SURVIVABILITY

Major Field: Industrial Engineering and Management

Biographical:

Education:

Completed the requirements for the Master of Science/Arts in Industrial Engineering and Management at Oklahoma State University, Stillwater, Oklahoma in July, 2015.

Completed the requirements for the Bachelor of Science/Arts in Production Engineering at Fr. Conceicao Rodrigues College of Engineering (CRCE), Mumbai, Maharashtra/ India in 2011.

Experience:

*Graduate Research Assistant*, Oklahoma State University, Stillwater, OK  
May'14-July'15

*Part time Student Supervisor*, Oklahoma State University Dining Services,  
Stillwater, OK  
Aug'13-May'13

*Junior Project Engineer*, Festo Controls, Mechatronic Motion Solutions, India  
Aug'11-Jan'12

*Intern*, Siemens Ltd, Process Planning Dept, India  
Sep'10-May'11

