Forecasting Financial Schedules for Cancer Patients

P.M.U.A.Jayatilaka

149214T

Faculty of Information Technology, University of Moratuwa, Sri Lanka

Forecasting Financial Schedules for Cancer Patients

P.M.U.A.Jayatilaka

149214T

Dissertation submitted to the Faculty of Information Technology, University of Moratuwa, Sri Lanka for the partial fulfillment of the requirements of the Degree of Master of Science in Information Technology.

May 2017

Declaration

I declare that this research is my own work and has not been submitted in any form for another degree or diploma at any university or other institution of teritary education. Information derived from the published or unpublished work of others has been acknowledged in the text and the list of references is given

Name of the Student

Signature of the student

P.M.U.A.Jayatilaka

Date:

Supervised by

Signature of the supervisor

Mr. S.C.Premarathne

Date:

Dedication

I dedicate this thesis to my parents who have always been my nearest and reverse nearest neighbors and have been so close to me that I found them with in me whenever I needed. It is their unconditional love that motivates me to set higher targets.

Acknowledgement

I would like to express my sincere gratitude to my supervisor Mr.S.C.Premarathne, Senior Lecturer in University of Moratuwa ,Sri Lanka whom expertise in the field of Information and Technology, encouraged me in understanding and applying knowledge in practice considerably throughout my research period. I do appreciate his vast knowledge and skills in many areas and his assistance and patience on me. Furthermore, I would like to thank the other lecturers of Department of Information Technology, University of Moratuwa especially Prof.Asoka Karunanada who taught us Research Methodology and Literature Review and Thesis writing subjects which were the basis for this research.

More over a very special thanks goes to Dr. B. I. Kurukulasuriya Deputy Director, National Cancer Institute Maharagama and his staff members for assisting me in collecting data sets. And also Dr. Y. Ariyarathne for granting me the permission to access his patients medical records , Dr.(Mrs.) K.K. Murthi for helping me to translating the medical records in to readable format.

I would also like to thank all the batch mates of MSc. in IT degree program who gave their valuable feed backs to improve the results of the research .Finally I would like to thank my family for the support they provide me through my entire life in particular.

Abstract

During the time a patient is admitted in hospital rich sources of clinical, bio medical, contextual, and environmental data about patients have been available in medical and health sciences. These clinical sources of information are marked increasing in both volume and variety. Due to continuous increasing of the size of health care data, certain complexity raised in it. The hidden patterns among patient data can be extracted by applying different techniques. The techniques and tools are very helpful as they provide health care professionals with significant knowledge towards a decision.

This research has been conducted to analyze cancer patient's medical records in Sri Lanka in an effective manner to predict the future cost estimation for medicine. It is hypothesized that analyzing cancer patient's medical record can be done through machine learning data mining techniques. By doing so we can help the patients to schedule their financial matters in coming years. In Sri Lanka almost 23,000 individuals diagnosed with a new invasive cancer each year. The costs of the diagnosis and treatment for the cancer are considerable important. These costs have increased over the past years and are expected to increase in the future. It has been claimed that most of the families face difficulties with a diagnosis of cancer have financial issues. To respond to this challenge, we developed an interactive application which utilizes a multiple linear regression model to forecast financial schedule for cancer patients

This solution take data set collected from Apeksha Hospital -Maharagama as the input and predict the factors associate with the question. Having received the input this approach prepossess the data set to remove the anomalies. Then build the data model to predict the future cost for the patient. Linear regression technique is used to develop for prediction of future cost estimation. This total solution is build using WEKA data mining software. (WEKA GUI and WEKA API). According to the evaluation of the predictive model the correlation coefficient value is 0.5725 and the root mean squared error is Rs. 553183.25.

Table of Content

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Prolegomena

During the time a patient is in hospital an extensive amount of data is collected for the hospital records. This includes ongoing progress notes, laboratory notes and results from any tests run including X-rays and electrocardiograms and operating reports from any surgeries. [1] Collected data in any hospital is now seen as a new source of very important information that can directly affect the efficiency of that hospital, provide higher quality outcomes, and even cut down the unnecessary expenditures.[2] Research in patient record analysis dates back early 1990s [3]. At present most of the hospitals in US has used patient record analysis for decision making in health care to improve the quality of service to the patients.[4] Despite numerous researches in patient record analysis has been done in overseas but not in Sri Lanka. This research presents our work on data mining approach to analyze cancer patient records.

1.2 Background and Motivation

One in three people will be diagnosed with cancer at some point in their lifetime. In SriLanka this translates into almost 22,000 individuals diagnosed with a new invasive cancer each year [5]. Altogether this means that there are increasing numbers of people living with cancer in Sri Lanka. The costs of the diagnosis and treatment for the cancer are considerable important. These costs have increased over the past years and are expected to increase in the future. There is growing awareness that cancer can have a major financial impact on newly diagnosed patients, those living with the disease, and their families. Indeed, it has been claimed that almost all families confronted with a diagnosis of cancer have financial issues or some kind of economic losses. [6] Most patients and families has to spend additional costs as a result of a cancer diagnosis. These can include direct medical costs such as seeing consultants, and those associated with buying drugs to help to cure the symptoms of cancer and the side-effects of

treatment. In addition, they have to spent more additional cost such as to traveling to hospital appointments. Increased utility bills are also common .

The most of cancer patients who are working need to take leave around diagnosis and during treatment and a most of them do not receive any sick pay from their employer. This may lead to decrease the household income of many patients and their families. By reduction in income with the additional costs has lead the cancer patients and their families to financial loss. Some have to borrow money from friends ,financial institutions or family or use their savings. On the other hand trying to reduce the household spending, and extra spending on clothes and leisure activities. Overall, this will increase the financial stress of the patient and their family members[7].

Medical Records usually contain five categories of information of patients. They are demographic information such as age, gender, address, race and ethnicity and other information of a patient, diagnostic information consist of disease names and severity of the diseases, laboratory indicators, where doctor orders includes drug name, delivery method of the drug, frequency and dose. Treatment of a patient is a series of doctor orders and an outcome of a patient can be cured, improve in effective, or dead [8] In the Sri Lankan context patient medical records are structured and formal that are given to patients to enable the continuity and quality of care which he takes with him when he goes for medical consultations [9].

In health care sector analysis of patient records has become increasingly popular because it offers lot of benefits to patients, hospitals, researchers, and insurance companies. By analyzing this data hospitals can identify effective treatments and best practices to the patients. By comparing causes, symptoms, treatments, and their adverse effects, data mining can analyze which courses of action are most effective for specific patient groups, help to identify clinical best practices. It makes better patient-related decisions and it leads to improve the patient satisfaction [10] .Hospital management is always under increasing financial pressure by analyzing patient records also influence costs, and operating efficiency while maintaining high-quality care. Patients can receive better health care services. Doctors can keep track of chronic diseases and identify high-risk patients, design appropriate action taken to improve a medical disorder, and reduce the number of hospital admissions and claims. Insurers can use this information to reduce their losses and the costs of health care.

EMR (Electronic medical records) have the potential to provide researchers with an incredible amount of information. Researchers have yet to model EMR complexity in a manageable way, and few people have investigated either its evaluation or established measures of performance. This has caused a problem differentiating the stakeholders' benefit of one electronic record over another [11].

While analyzing the medical records we can also predict there admissions of discharged patient in a short window of time. Many reasons could lead to re admissions of the patients such as early discharge of patients, improper discharge planning, and poor care transitions [12] . Marquardt [13] others used linear regression , regression trees to predict costs for the upcoming year, interactively select from a set of possible medical conditions, what are the factors that associated to the cost evaluated through by compare costs against historical averages.Meadem & others [14] treat the problem of readmission as supervised learning matter. Social ,demographic factors, health conditions ,disease parameters, hospital care quality parameters, and a variety of variables specific to health care providers were treated when build the predictive model .Logistic Regression ,Naive Bayes, Support Vector Machines use to build the predictive model .

According to Cao [15] health costs are increasing rapidly worldwide. Finding ways to reduce medical costs and allow more people to have access to medical care has become a huge problem that all countries are attempting to solve.Bhuvan [16] studied on Machine learning methods to identify diabetic patient who are having high risk of readmissions in the future (short term-before 30 days & long term-after 30days). Number of inpatient visits, discharge disposition and admission type were identified as strong predictors several classification models (Naive Bayes, Bayesian Networks, Random Forest, Adaboost and Neural Networks) were used in this study. Feature Analysis done by using Ablation Study of Risk Factors and associative Rule Mining(Apriori algorihm)

Due to continuous increasing the size of health care data a type of complexity is exist in it. It is very difficult to extract the meaningful information from the medical records while using traditional methods. Due to advancement in mathematical and

statistical fields ,now it is possible to use data mining to extract the meaningful patterns from it. Due to its approaches and boundless applications to mine the data in proper manner, data mining is becoming popularity in research fields. Data mining have a great impact on health care systems to use data more efficiently and effectively[I0] . Data mining is becoming popularity in research fields because to its approaches and boundless applications to mine the data in an proper manner. By using data mining to uncover previously hidden patterns from vast amount of data stores
and then use this data for building predictive models. In health care applications of data mining, performs better than the traditional methods, special characteristics of health data, and new health condition mysteries have made data mining very necessary for health data analysis [1] . However, data mining models and specifically predictive models, can be very helpful when they are used as a second opinion for the physician's decision on a treatment. with machine learning techniques and medical data involved numerical information, tests results, disease, diagnosis and treatment level status and other health care services information [17]. Data mining and machine learning are becoming the most interesting research areas and increasingly popular in health domain [18] Machine-learning techniques are providing magnificent work in various fields. Medicine is one of the fields that can benefit from the application of data mining and pattern recognition techniques [19]Therefore, data mining has received a lot of attention due to its strong ability of extracting information from data.

1.3 Problem statement

According to literature many of the analysis has done using the patient records to support for cost-savings and decision making purposes. Such as diagnosis and treatment that can be used by to doctor's medical decision support, predict the treatment plan to patients when there is no dis-positive evidence favoring a particular treatment option. In the health care resource management can help by predicting the patient length of stay to properly manage the resource allocation by identifying high risk patients and predicting the usage or need of various resources.
Customer relationship management can be helped by the identifying the usage and purchase patterns to improve the overall customer satisfaction. Fraud and anomaly detection also be help by analyzing to find prescription fraud and detection and prediction of faults in medical devices.

In Sri Lanka very few research has been carried out to analyze the patient records. Most of the solutions were built using statistical analysis methods. And the collected data was based on the closed ended questions.

During this study I have identified there were no adequate research carried out in Sri Lanka to identify patterns attach with patient records. Specially to analyze future cost estimation for medicine. In order to solve that problem I propose a data mining approach to analyze the patient records & predict the outcomes of that analysis.

1.4 Hypothesis

There were no adequate research carried out in Sri Lanka to identify patterns attach with patient records. In order to solve that problem I propose a data mining approach to analyze the patient records for financial forecasting.

1.5 Objectives

(i) To critically review the state of the art of patient record analysis

(ii) To do an in depth study of data mining in patient record analysis, algorithms with a particular emphasis on machine learning algorithms

(iii) Develop a data cleaning algorithm to cleans the data set (by removing unwanted date and filling missing values)

(iv) Feature selection and model building

(v) Predict the cost estimations for the future based on the previous records

1.6 Data mining-based approach for analyze cancer patient records.

In this research the data set required was obtained with the assistance of the "Apeksha Hospital – Maharagama. All the details in the patients record card such as age, gender, disease , number of re admissions, side effects , number of medications etc. has been considered when developing the predictive model.

Analyzing patient records with data mining all the standard steps in knowledge discovery process which includes learning the application domain, data selection, data cleaning and pre processing, data reduction and projection, data mining and analysis of results, visualization, transformation, removing redundant patterns, etc. the collected These records may have noise data because some of these records have some missing values and some have form filling errors (human errors) as well. (because all these records are paper based), so the noise data need to be eliminated. Then data is transformed in to forms appropriate for mining by performing aggregation and summary.

Association and classification methods used to extract the patterns to discover the hidden knowledge.Machine learning algorithms were used in the prediction model And the results obtained from the mining data would be available in the hard copy printed form and the output could be in Graphical manner such as images, graphs (line charts).

1.7 Structure of the Thesis

The rest of the thesis is organized as follows. Chapter 2 critically reviews the literature on patient record analysis and identify the research problems. Chapter 3 is about the technology used for analysis of patient records. Chapter 4 present our new approach to use to predict there admissions and future cost estimation. Chapter 5 and Chapter 6 describe the design and implementation respectively. Chapter 7 is on evaluation of the new solution. Chapter 8 concludes the research with a note on further work.

1.8 Summary

This chapter gave an overall picture of the entire project presented in this thesis. As such we described the background/motivation, problem definition, hypothesis, objectives, and a brief overview of the solution. Next presents a critical review of literature on patient record analysis.

Chapter 2

Developments and Challenges in Patient Record Analysis

2.1 Introduction

Chapter 1 gave a comprehensive description of the overall project described in this thesis. This chapter provides a critical review of the literature in relation to developments and challenges in Data mining. For this purpose the review of the past research have been presented under three major sections, namely early developments, modern trends and future challenges. At the end of this chapter define the research problem as the no adequate research carried out in Sri Lanka to identify patterns attach with patient records and identify the data mining machine learning technology that can be used to address the problem.

2.2 Early Developments

Authors [11] used classification trees and clustering algorithms, and medical claims data to provide predictions of health-care costs in the third year by applying data-mining methods to data from the first two years. Based on the reviews [20] used to calculate the future health care cost many statistical models have been used ,such as R-squared, predictive ratios, prospective models, combined (i.e. prospective and retrospective) models. Moturu and others [21] studied on demographic and disease-related (inpatient information, pharmacy information) features to identify the future high-cost patients. support vector machine classifier, logistic regression, logistic model trees, ada boost and logit boost were used as learning model. Sensitivity, specificity and f-measure used for check to validate the model . Study was carried out by Jasti & others [22] to identify the risk factors associate with re admission of patients . From that they have identified major 42 risk factors are associate with the re admissions.

2.3 Modern Trends in Patient Record Analysis

During the health care treatments vast amount of medical data is collected in the patients medical record. Challenging task is to analysis of these records since a vase amount of hidden knowledge can be automatically mined to effectively. By doing so this will help both physicians and health care organizations [23] . Several big data analytic shave been done for patient record analysis. Out of that Herland [24] has shown the impotence in big data analytic in health care. Roosan [25] study in big data based decision support systems .Zolfaghar and others [26] studied big data solution for readmission of patients and done the predicting of readmission using data mining model. They use HIVE & Cassandra to extract & integrate patient data set. Prediction done using Mahout distributed analytic platform.

They selected random forest algorithm as the classification algorithm. Problem is treated as binary classification task (in supervised learning).Social demographic, vital signs, laboratory tests, discharge disposition, medical comorbidity and other cost related factors, like length of stay used as predictor variables. Ojha & others [27] proposed method to analyze patient record using big data solution, by Hadoop and HDFS (Hadoop file system) tools. Vaishali [28] & others implement solution using Map Reduce technique with apache hadoop. Chandiok [29] used Machine Learning (Decision Tree) and Fuzzy-Cognitive-Map Model (Hebbian learning) for Cognitive medical decision making systems.According to [30] Multi-agent architectures were also used in clinical decision support Systems .

In the same token many researchers have done several studies under this problem out of that Hao wang and others [31] have implemented a solution using LACE index for re admissions. The LACE index score was calculated on each patient using length of hospitalization stay ("L"), acuity of the admission (''A"), comorbidities of patients ("C''), and emergency department use of patients (''E"). Variables uses in this work are basic clinical characteristics, length of hospitalization, comorbidities, number of previous emergency department visits, and the number of post discharge emergency department visits for 30, 60, and 90 days.

Logistic regression, decision tree & neural network use to build the models for predict the readmission of 28 days mis classification rate and root ASE use to evaluate the predictive power of the three models as well as the lift chart and ROC curve, were

also used. For this study, risk factors used in the demographic category were sex, age, and region of residence for the treatment and clinical category were considered [32]

Kang and others [33] develop four data mining models including two artificial neural network models and two classification and regression tree models to predict the hospital charges and amount paid by the insurance for cancer patients. Neural network model used feed-forward back propagation method, decision tree model, RELIEFF methods. Clementine 7.0 use to build the models RELIEFF method use as feature selection

Chechulin and co-workers [34] carried out research using logistic regression model to predicting the high cost healthcare patient in the future. Information was collected on various demographic and utilization characteristics (54 variables were treated) were considered when building the model. Performance of the model was evaluated using C-statistic for predictive ability of the model.

Authors [35] have implemented model for predicting the Pneumonia Readmission using the RBF-SVM. Demography variables (gender and age), treatment and clinical variables (number of comorbidities, number of treatment procedure, number of medication), Elixhauser comorbidity measure, health care utilization factors (length of admission and total cost of admission), and biochemistry examination variables (blood BUN, creatinine, albumin, sodium, glucose,C-reactive protein, procalcitonin, white blood cell, neutrophil, lymphocyte, hemoglobin, and platelet), were considered when building the model. Out of that six significant variables (age, gender, number of medication, length of admission, number of comorbidities, and total admission cost) were mostly considered

Sun and others [8] have carried out research on recommending treatment plans from patient record analysis using density Peaks based Clustering (DPC) algorithm ,where as it can discover clusters with complex shapes, while traditional exemplar-based clustering algorithms can only find spherical clusters

Conrad and co-workers [20] studied on Unsupervised neural network(self-organizing Maps -SOM) solution for feature extraction from patient discharge summaries. Vector space model were used for pre processing stage with parsing and indexing.

Zhou and co-workers [36] used text numng approach usmg natural language processing techniques for analysis of patient records. Also Champion [37] proposed solution to Improved Patient Characterization ,using text mining solution with natural language processing techniques. In the same token Ray [38] has conduct health care data analytic using text similarity-based mining.

Bruno & others [39] have proposed method using data mining approach, based on a density-based multilevel clustering algorithm to identify the examinations commonly followed by patients with a given disease. To classify the clusters they used decision tree classifiers .The following attributes were considered when they creating the decision tree: patient age, gender, and the examinations done by patients weighted through the TF-IDF weighting score. Accuracy, precision and recall methods were used to evaluate the quality of constructed classification model

Zolfaghar and coworkers [40] proposed model for predicting risk-of-readmission as a supervised learning problem, using a multi-layer classification model. They split the problem in to 3 stages namely ,a)at risk in general (b) risk within 60 days (c) risk within 30 days. This is the main difference from the other models which they try to predict the readmission by direct model. From this model they have used different classification models which are suitable for each stage. They have used chi-square test as the feature selection technique in each layer. 50 predictor variables used in this model including age, gender,Blood Pressure,Ejection fraction value,etc.They used Naive Bayes & Support Vector Machine as evaluation methods.

Zolfaghar and coworkers [41] develop online tool to analyze and predict clinical risk for 30 day risk of readmission using in compute patient data as input. Clustering based approach has been used to solve the problem.

Ximeng Liu & others [42] have studied privacy-preserving patient centric clinical decision support system on Naive Bayesian Classification. Authors have used naive bayesian classifier to compute the disease risk of new coming patient in secure manner , and results were evaluated via extensive simulation. They have also implement photographic tool called additive homomorphic proxy aggregation scheme to preserve the privacy of the cloud data.

Study carried out by [43] to identify the clinical and laboratory markers associated with hospital readmission in heart failures using multi variable logistic regression analysis.

Zikos [44] and others were proposed method to predict the hospital length of stay using machine learning techniques including Naive Bayes, Ada Boost and C4.5 Decision tree, for two different LOS cut-off points (4 day and 12 day hospital stay). Patient demographics, admission information were considered According to authors [45] Support Vector Machines (SVM), Decision Trees (DT) and Naive Bayes (NB) techniques was used to predict the readmission in ICU. Authors [18] used four different classification methods for prediction model: Decision tree, Naive Bayes, SVM, and ANN and to reduce classification error they used adaptive boosting and stacking. Clustering based approach (Self Organizing Map) and Genetic K-Means algorithm were used by to explore health care data

Genetic algorithm (used as a feature selection technique)and random forest (used as a classifier) used for diagnosing lymphatic diseases by Elshazly [19]

Ming Zhou [36] used machine learning, statistical feature selection and genetic algorithms to identify early predictors of disease progression about ankylosing spondilitis illness phenotype.

Chatterjee & others [46] studied on Artificial Neural Networks based model to predict blood sugar level based on previous day's record (records were collected daily using text messages)

Study was carried out by [47] and others to find there admissions, 45factors including patient history findings, physical examination findings, laboratory findings and chest x-ray finding were considered to develop the model. Logistic regression, Random forest, logit boost used as learning methods for readmission.

In Sri Lanka very few research has been carried out to analyze the patient records, out of that Dissanayake & others [48] have conduct research to assess the frequency of diseases using statistical analysis method. Rathnayake & co-workers [49] done study on privacy, confidentiality and security issues pertaining to electronic medical records in Sri Lanka. In the same token Riyaz [50] has done study on factors associate with in

the re admissions of elderly people in Sri Lanka using statistical analyze technique by giving them a structured closed type question.

2.4 Future challenges

According to [51] problem of predicting heart disease, severity of diabetes, dengue symptoms, malaria symptoms has been a major challenge for the researchers from the medical domain as well as data mining and information retrieval domain.

2.5 **Summary**

This chapter presented a comprehensive literature review on the patient record analysis and identified the research problem as the inadequate attention given to analyze the patient record in Sri Lanka . We also identified data mining technology to address the above problem. Next chapter will discuss the technology to be used for our solution.

Chapter 3

Technology Adapted - Data mining Tools and Techniques

3.1 Introduction

Chapter 2 discussed the existing methods and attributes for analyzing patient records for predicting the re admissions and cost associate with it. This chapter presents data mining technology which is selected to analyze the patient's records effectively in detail. This chapter highlights the effectiveness of selected technology that distinguish it from the technologies applied in existing literature.

3.2 What is Data mining?

There is a vast amount of data available in the Industry of Information . Until this data is converted in to useful information this data is of no use. It is necessary to analyze this data and extract useful and hidden information from it. Data mining is the process of automatically discovering useful information in large data repositories. Data mining technique are deployed to scour large databases in order to find useful patterns associated with it.

The overall process of finding useful knowledge in raw data involves sequential line up of steps such as developing and understanding of the application domain, creating a target data set based on an intelligent way of selecting data, data cleaning, data integration, data transformation, Data Mining, pattern evaluation and data presentation[52] . Data mining process is shown in Figure -3.1 .

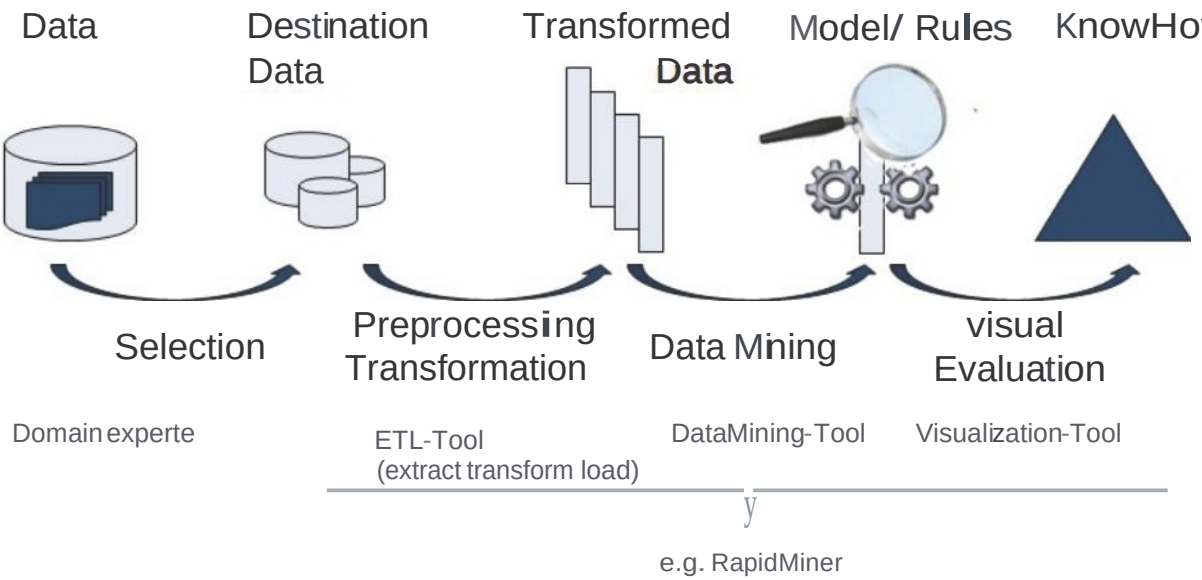

Figure -3.1- Steps of the data mining process

3.3 Data Mining Techniques

Data mining tasks are generally divide in to two sections such as predictive tasks (which predict the value of a particular attribute based on the values of other attributes) and descriptive tasks (which derive the patterns that summarize the underlying relationship in data). Techniques such as classification, regression and time series analysis are mostly common used predictive tasks, whereas association rules, cluster analysis are used for descriptive tasks. There for selecting the most suitable mining techniques depends on the problem that user going to address. Some of the data mining techniques are shown in Figure -3.2 .

In this research the problem is to find the future cost for that we need to select a regression technique. In a regression problem, we are trying to predict results within a continuous output, meaning that we are trying to map input variables to some continuous function.

Figure -3.2- Data mining techniques

3.3.1 Multiple Linear Regression

Linear regression is a common Statistical Data Analysis technique. [t is used to determine the extent to which there is a linear relationship between a dependent variable and one or more independent variables. Linear regression is divided in to two types namely simple linear regression and multiple linear regression.

[n simple linear regression a single independent variable is used to predict the value of a dependent variable. Two or more independent variables are used to predict the value of a dependent variable in multiple linear regression. The difference between the two is the number of independent variables. [n both cases there is only a single dependent variable.

The dependent variable must be measured on a continuous measurement scale (e.g. 0-100 price) and the independent variable(s) can be measured on either a categorical

(e.g. male versus female) or continuous measurement scale. There are several other assumptions that the data must satisfy in order to qualify for linear regression. When selecting the model for the analysis, another important consideration is the model fitting. Adding independent variables to a linear regression model will always increase the explained variance of the model (typically expressed as R^2). Adding more and more variables to the model makes it inefficient and over fitting can occur. The other concern of regression analysis is under fitting. This means that the regression analysis' estimates are biased. Under fitting occurs when including an additional independent variable in the model will reduce the effect strength of the independent variable(s). Mostly under fitting happens when linear regression is used to prove a cause-effect relationship that is not there. [55]

3.3.2 SMO Regression

SMOreg implements the support vector machine for regression. The parameters can be learned using various algorithms. The algorithm is selected by setting the RegOptimizer. This implementation globally replaces all missing values and transforms nominal attributes into binary ones. It also normalizes all attributes by default. (Note that the coefficients in the output are based on the normalized/standardized data, not the original data.) [56]

3.4 Data Mining Tools

There are many ready made tools available for data mining today. Some of these have common functionalities packaged within, with provisions to add-on functionality by supporting building of business-specific analysis and intelligence [53].

3.4.1 WEKA Data Mining Tool

WEKA (Waikato Environment for Knowledge Analysis) is a JAVA based customization tool,and it is very sophisticated and used in many different applications including visualization and algorithms for data analysis and predictive modeling techniques such as clustering, association, regression and classification. Its free under the GNU General Public License .

3.5 Java

The prime reason behind creation of Java was to bring portability and security feature into a computer language. Java is platform independent language.The Java platform differs from most other platforms in the sense that it is a software-based platform that runs on the top of other hardware-based platforms. Java multithreading feature makes it possible to write program that can do many tasks simultaneously. Also WEKA API is developed using java language, so using java for accessing the WEKA API, is most suitable than using other languages [59].

3.6 Summary

This chapter presented data mining as the technology proposed to analyze patient records to predicting the future cost for medicine. In this sense, it is pointed out how the data mining offers an efficient and accurate solution for patient records analysis. The next chapter shows the approach of analyzing cancer patient records through the technology presented here.

Chapter 4

Approach to Forecasting Financial Schedules For Cancer Patients

4.1 Introduction

Chapter 3 discussed the technology for analyzing the patient records. This chapter present our approach to analyze patient records in detail using data mining under several headings namely, hypothesis, input, output, process, users and features. This chapter describes the selected approach for patient record analysis. Here we describe our approach on Forecasting Financial Schedules For Cancer Patients, FFSCP.

4.2 Hypothesis

There were no adequate research carried out in Sri Lanka to identify patterns attach with patient records. In order to solve that problem I propose a data mining approach to analyze the cancer patient records for forecasting financial schedules for cancer patients.

4.3 Users

Number of users who can be benefited by the FFSCP systems in multiple ways. More importantly health care providers, patients, and insurance companies can be directly benefited by this solution. Those who are interested (eg: researchers) in patients record analysis can also use this system for learning purpose.

4.4 Input

All the details in the patients record card such as age, gender, medical specialty, current stage , side effects , number of medications etc.are taken as input to build the model. (variables of the collected data set is attached in the Appendix -A)

4.5 Out put

The output of the system would be available in the hard copy printed form and the output could be in graphical manner such as images, graphs (line charts)

4.6 Process

In this process of analyzing patient records with data mining all the standard steps in knowledge discovery process (Figure 4.1) which includes learning the application domain: relevant prior knowledge and goals of application, data selection - creating a target data set, data cleaning and pre processing, data reduction and projection, find useful features, dimension/variable reduction, invariant representation, choosing the mining algorithm(s), data mining: search for patterns of interest, interpretation: analysis of results, visualization, transformation, removing redundant patterns, etc .

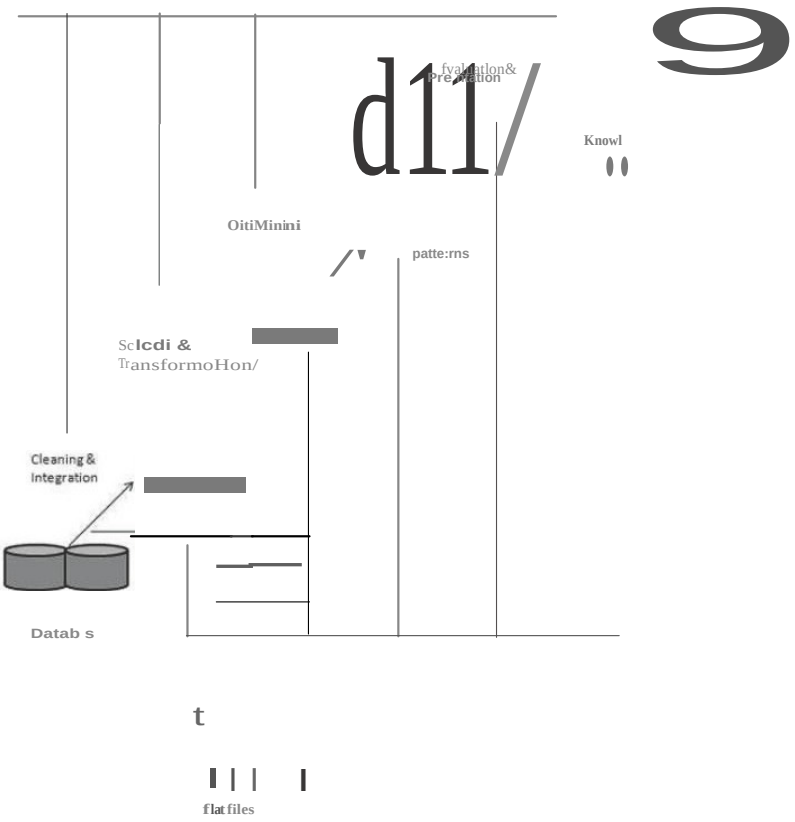

Figure 4.1 - Steps in knowledge discovery process

4.6.1 Data Selection

To determine the future cost for medicine, we need to keep track of the basic details in the patient record .In the Sri Lankan context patient medical records are formal and structured records, which includes patient demographics details progress notes, vital signs, medical histories, diagnoses, medications, immunization, allergies, radiology images, lab and test results (Administrative and billing data were also considered in private hospitals).

In this research the data set required was obtained with the assistance of the "Apeksha Hospital -Maharagama" . This material (all the record are in written format) includes 507 patient records from ward 07 and 08 patient who are suffering from Leukemia, Breast cancer and Thyroid cancer from year 2010,2011,2012. (Death patient's and pediatric records during the period were not considered) , initially the cost details are not included in the record set there for cost values are filled with approximation values for the medication costs.

4.6.2 Data Pre processing

Real world clinical data is noisy and heterogeneous in nature, severely skewed, and contains large number of pertinent attributes. For this research data is collected for Apeksha Hospital -Maharagama. These records may have noise data because some of these records have some missing values and some have form filling errors (human errors) as well. (because all these records are paper based). So the noise data need to be eliminated. Accuracy, completeness, consistency, timeliness, be-livability, value added ,accessibility and interpretability are some of the characteristics counted on when data taken to a research to draw well accepted conclusion. There for these data set is pre processed before further analysis

4.6.3 Data Transformation

In this step data is transformed in to forms appropriate for mining by performing aggregation and summary. Due to the huge volume of data , tuning those data to useful knowledge to support decision making we use smoothing for remove noise data, aggregation for summarization, data cube construction and generalization for concept hierarchy climbing and normalization is used to scale with in small specified range. According to our problem appropriate smoothing and aggregations are done.

4.6.4 Data mining

This step is the most important part of the whole research . We need to use appropriate methods to extract the patterns to discover the hidden knowledge. Association and classification methods has to be consider in mining the data. In the association, the relationship of given item in a data transaction on other items in the same transactions are used to predict the patterns. In classification methods are intended for learning different functions that map each item of the selected data in to one of the predefined set of classes. According to our task predictive mining task is selected.

Regression and classification are both related to prediction, where regression predicts a value from a continuous set, whereas classification predicts the belonging to the class.

4.6.4.1 Data mining -Classification vs Regression

Predict the target class (Yes/ No) is known as classification. If the trained model is for predicting any of two target classes. It is known as binary classification. For example consider the patient is a high cost patient or not? These kind of problems will be addressed with binary classification.

To predict the discrete or a continues values , regression algorithms are used. In some cases, the predicted value can be used to identify the linear relationship between the attributes. Eg. to predict the cost for patient, Based on the problem difference regression algorithms can be used. Some of the basic regression algorithms are linear regression, polynomial regression, etc

For this research problem we need to find out the cost for future medicine for financial prediction purpose. There for we need to use regression analysis not for classification.

4.6.5 Evaluation/interpretation

In this step we show the results obtained from the mining data. In order to properly interpret the knowledge patterns, it's important to use an appropriate visualization tool.

This solution is works on windows 7 operating system and uses WEKA as the data mining software. WEKA, is an open source data science platform, visually-based

software accelerates the process of creating predictive analytic models and makes it easy to get the results embedded in business operations.

4.7 Features

The solution proposed by this research can be used to analyze huge volume of cancer patient record data, through the predictive mining tasks this solution allows to make prediction for future instances. This solution provides the output by extracting previously unknown patterns dynamically.

4.8 Summary

This chapter presented our approach to analyze the Sri Lankan cancer patient records to predict the future cost estimation for medicine. It is pointed out how the approach offers an efficient and accurate solution using machine learning algorithms in data mining. The next chapter shows the design of the approach presented here.

Chapter 5

Design of FFSCP

5.1 Introduction

The previous chapter gave **full** picture of the entire solution. This chapter describe the design of the process presented in the approach. Here we describe the top level architecture of the design by elaborating the role of each component of the architecture.

5.2 Top level Architecture of FFSCP

This shows the basic structure of the design. A user can enter data with details like age, gender, diagnosis ,stage and treatment through a user interface. The data submitted by the user is written to .arff file (attribute relation file format) then file is sent to the WEKA API. The API parses the data into the predictive model. The predicted values are sent back through the WEKA API to the user. Figure- 5.1 illustrate the top level architecture of the FFSCP.

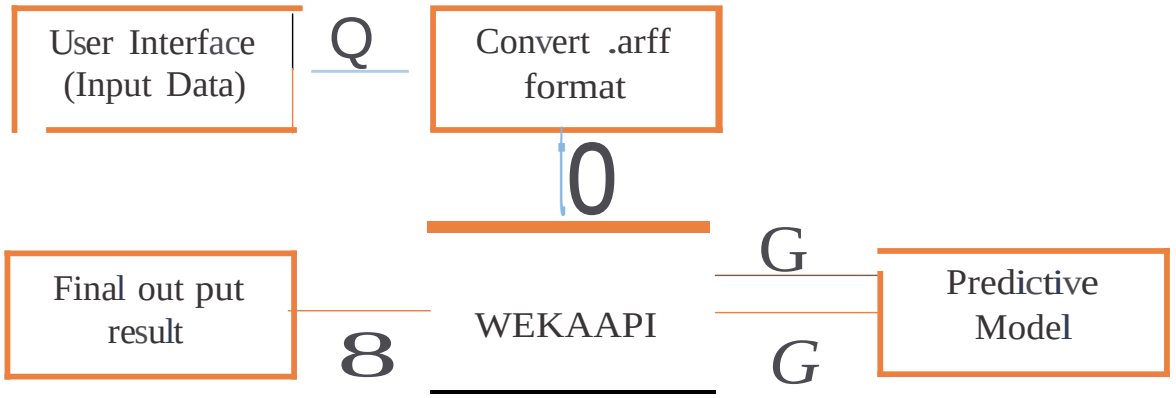

Figure- 5.1 - Top level architecture of FFSCP

5.3 Data Model of FFSCP

Figure- 5.2 illustrate the data model of the FFSCP. This shows how the data set is divided for training the model and testing the build model.

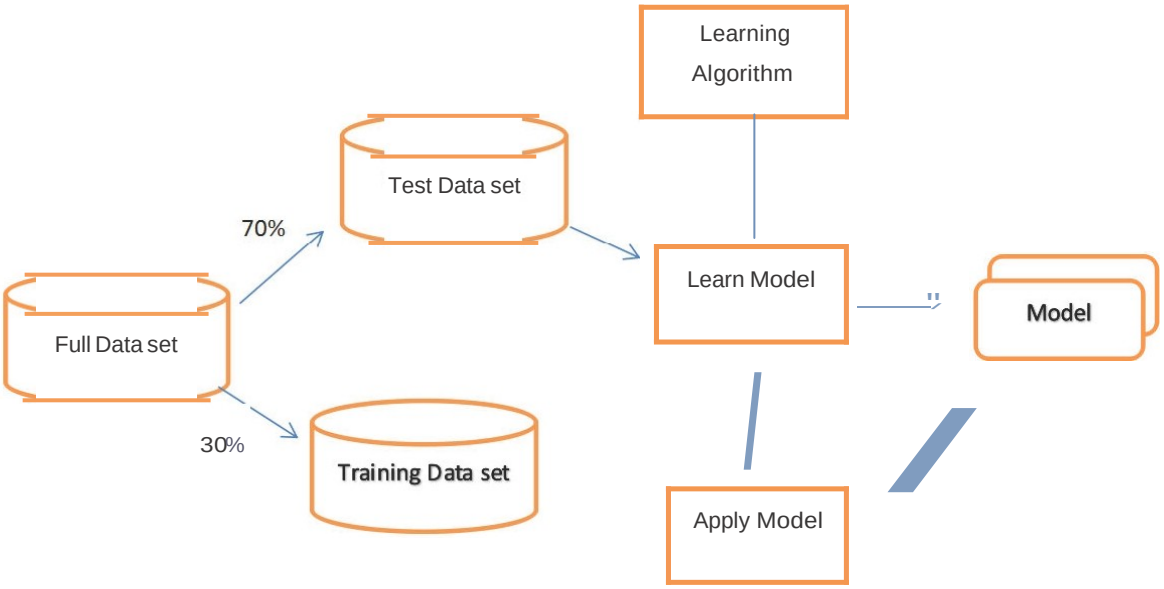

Figure- 5.2- Data model of FFSCP

5.4 Cost Estimation Module

This module illustrate what are the factors to be consider when predicting the future cost of a patient. This will follow several steps namely extraction, preprocessing , prediction and evaluation as shown in Figure 5.3.

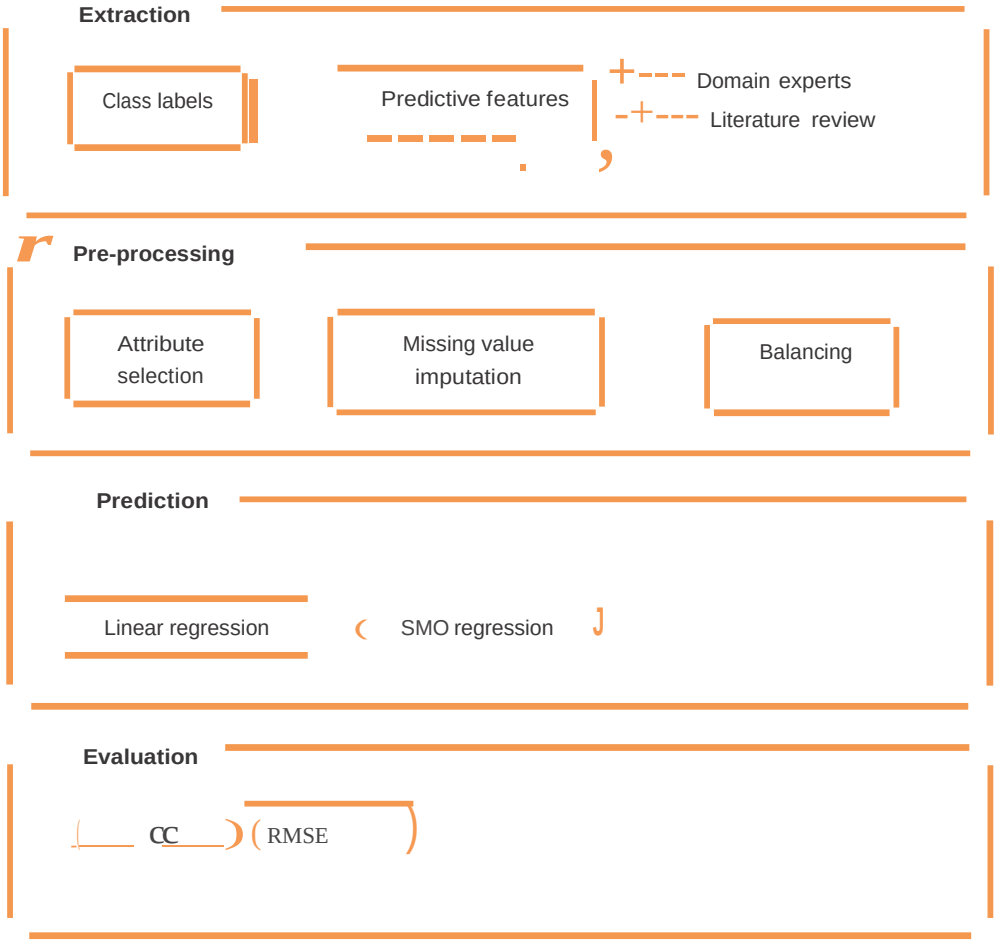

Figure 5.3- Cost estimation module

CC-Correlation coefficient , RMSE -Root mean squared error

5.5 User Interfaces

User interface offers facilitates to interact with the system for users. It retrieves patients data as the input details and suggest the future financial cost for patients based on the future treatments. Figure 5.4 illustrates the main interface.

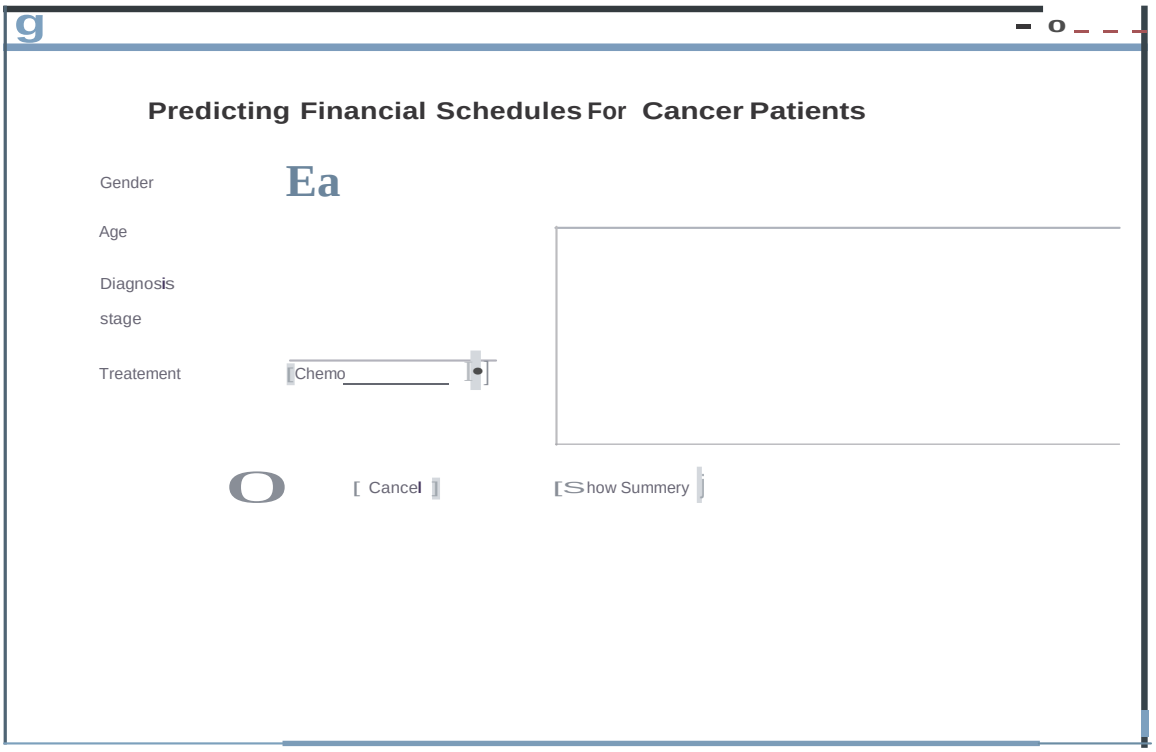

Figure 5.4 -User Interface

5.6 Summary

This chapter provided details on research design and applicability of selected research method for the research. Further more this chapter focuses on top level design for the research and how research question are structured with in the research. Subsequent section will be discussed about implementation details according to this design.

Chapter 6

Implementation of FFSCP

6.1 Introduction

In chapter 5 the top level design of the solution has been described in terms of what attributes are used to represent patient record analysis. This chapter describes the implementation of the problem regarding software, algorithms, method ,etc.

6.2WEKA

Waikato Environment for Knowledge Analysis *(W*eka) is a suite of machine learning software written in Java. Weka contains a collection of visualization tools and algorithms for data analysis and predictive modeling, together with graphical user interfaces for easy access to these functions. Weka supports several standard data mining tasks, more specifically, data
preprocessing, clustering, classification, regression, visualization, and feature selection. All of Weka's techniques are predicated on the assumption that the data is available as one flat file or relation, where each data point is described by a fixed number of attributes.[54]

6.3 Data Collection

Data for this analysis is collected form Apeksha Hospital Maharagama. All the diagnosis data and treatment data are taken from the patients records and manually calculated the treatments cost according to the treatments. A sample attributes of the data set is attached in Appendix -A.

6.4 Pre-processing of the Dataset

Once the initial domain knowledge has been amassed and initial important factors pertinent to the problem are identified, the next task ahead is to pre-process the data to make it amenable for building predictive models.

6.4.1 Missing Valoe Imputation

Collected data had some of missing values and noisy values. We have some missing values for the variable "stage", we manually removed (11 records) those entries because those records do not have any follow up on the treatments. We followed the process shown in Figure 6.1 to pre process data fill in missing values is done using Weka's filter "Replace Missing Values" which replaces all missing values for nominal and numeric attributes in the data set with the modes and means from the training data. The class is ignored while applying it.

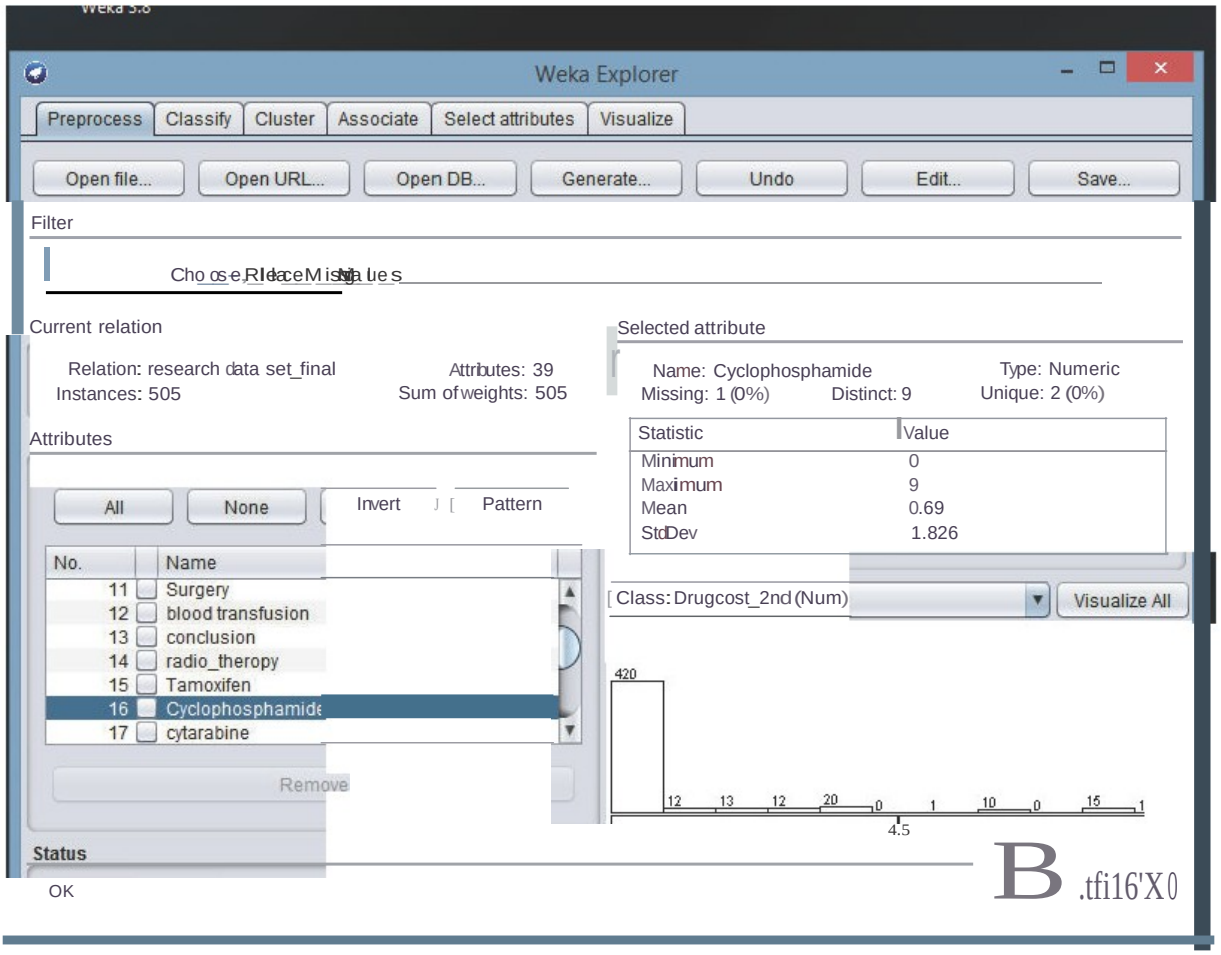

Figure 6.1-Replace missing values

6.4.2 Feature Selection

Because of the complexity and uniqueness of the domain, hospital admission due to cancer is a complex phenomenon governed by multiple features. One of our major challenges before the prediction task is to determine the subset of attributes that have

a significant impact on cost of patients from the attributes present in the data set. Before we apply the feature selection we need to apply a filter for turning numeric attributes into nominal ones. For this purpose we apply NumericToNominal Filter (Some screen shots which shows in the Figure 6.4.1 Appendix B). Then we consider state-of-the-art feature selection technique: ReliefFAttributeEval which evaluates the worth of an attribute by repeatedly sampling an instance and considering the value of the given attribute for the nearest instance of the same and different class and Ranker search method.(Ranks attributes by their individual evaluations) Figure 6.2 Shows the selected attributes based on the Rank. Applying the attribute selection filter to select only necessary attributes with the evaluator ReliefFAttributeEval)
Then all numeric values are converted in to binary values using NominalToBinary filter. (Some screen shots which shows in the Figure 6.4.2 in Appendix B.)

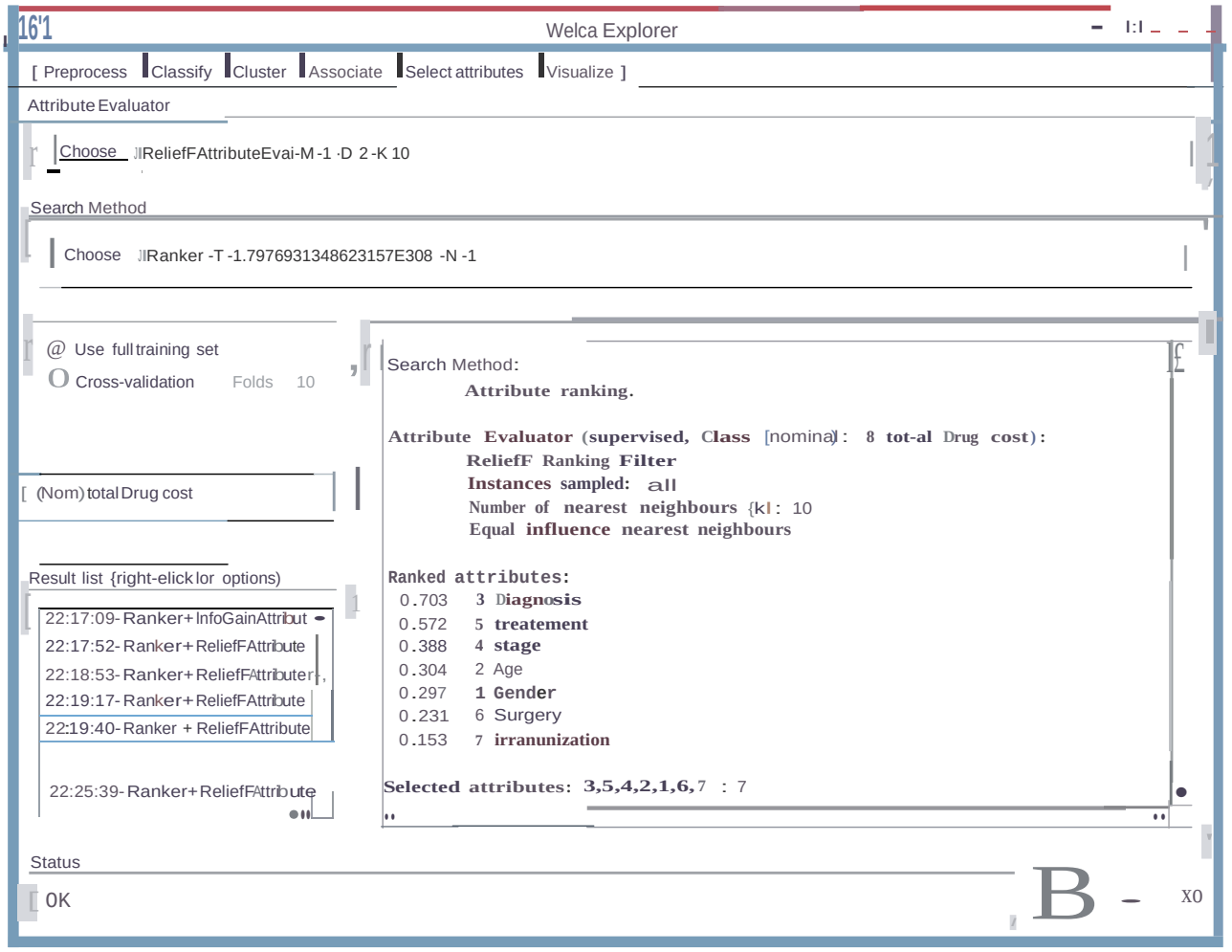

Figure 6.2- Attribute selection

6.5 Data Model using WEKA

To build the predictive models we use WEKA software.

6.5.1 Regression Models for Future Cost Estimation

First we build several models for predicting the cost , and based on the results we are going to select the best model which fits for the problem.

6.5.1.1 Multiple Linear Regression

Linear regression is a common statistical data analysis technique. It is used to determine the extent to which there is a linear relationship between a dependent variable and one or more independent variables. In this research we use multiple linear regression which takes two or more independent variables are used to predict the value of a dependent variable.

In here we select the dependent variable as total cost , and independent variables as age, diagnosis,gender,stage and treatment used by the patients. We split the data set as70% for train the model and 30% for test the model. Figure 6.3 in Appendix - B shows the build model for the analysis.

6.5.1.2 SMO Regression

SMOreg implements the support vector machine for regression. This implementation globally replaces all missing values and transforms nominal attributes into binary ones. It also normalizes all attributes by default. We use filter type as Normalized training data and select poly kernel as kernel and used RegSMOimproved as regoptimizer.

In here we select the dependent variable as Total cost , and independent variables as age, diagnosis,gender,stage & treatment used by the patients. We split the data set as70% for train the model and 30% for test the model. Figure 6.4 in Appendix B shows the build model for the analysis.

6.6 Implementation of the Application

Application is implemented using java and WEKA API. Patient can insert basic details (gender, age, Diagnosis, stage,treatment) of the cancer then the application will suggest the next treatments & cost for the future medicine. Graphical view is also available in the application. Sample is shown in Figure 6.5. Figure 6.6 in Appendix – B shows the graphical view of the results and sample codes of the application are attached in Appendix -C.

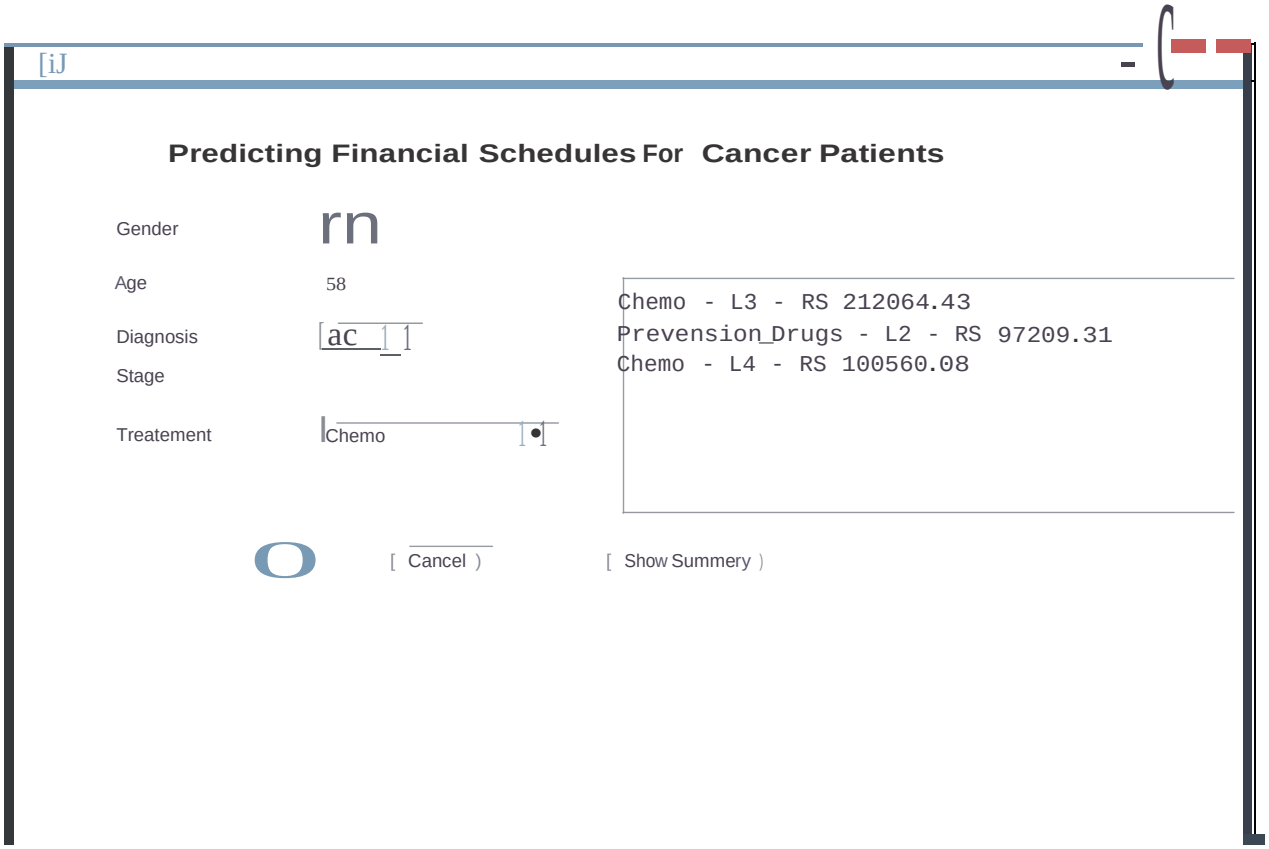

Figure 6.5 -Application Interface with results

6.7 Summary

Implementation chapter provide the full path in constructing data model for addressing the research questions. Furthermore this chapter gives detail description about using WEKA to build the model Next chapter will be on discussion about evaluation.

Chapter 7

Evaluation

7.1 Introduction

This chapter focuses on how testing strategies carried out for the research question in terms of the evaluation measurements for the selected data mining techniques. The system is tested in terms of users, and how the selected model tested by using test data set.

7.2 Data Model Testing on Regression Models

To evaluate the accuracy of the build model we need to check the correlation coefficient and root mean squared error of the model.

7.2.1 Correlation Coefficient (CC)

This measures the strength and the direction of a linear relationship between multiple variables. The value of CC ranges between from **-1** to 1.Values of CC close to 0 imply that there is little to no linear relationship between the data. Values of CC close to **1** shows that there is a positive linear relationship between the data. Values of CC close to -1 shows that there is a negative linear relationship between the data. [57]

7.2.2 Root Mean Squared Error

The RMSE (Root mean squared error) is the square root of the variance of the residuals. It indicates the absolute fit of the model to the data-how close the observed data points are to the model's predicted values. RMSE is an absolute measure of fit. Lower values of RMSE indicate better fit. RMSE is a good measure of how accurately the model predicts the response, and is the most important criterion for fit if the main purpose of the model is prediction. [58]

7.3 Data model Evaluation

Wetest the build data model with multiple linear regression on 30% of our initial data set. It shows correlation coefficient of 0.5725 and root mean squared error as 553183.2528. **(Fi re** 7.1-Appendix **-B)**

SMO (support vector machine) regression model tested on 30% of our initial data set. It shows correlation coefficient of 0.5889 and root mean squared error as 581000.9477. **(Fi re** 7.2- Appendix-B)

Summary of the evaluation results are shown in Table 7.1

	SMO Regression	Linear Regression
Correlation coefficient	0.5889	0.5725
Mean absolute error	223944.6529	333645.4507
Root mean squared error	581000.9477	553183.2528
Relative absolute error	43.6175 %	64.9838 %
Root relative squared error	86.068 %	81.9472 %

Table 7.1 - Sununary of the evaluation results

According to the results of the evaluation we selected the multiple linear regression to build the data model because it's having low root mean squared en·or (RMSE)than theSMO regression as well as In both cases correlation coefficient (CC) is not much different.

7.4 Swnmary

This chapter concludes with test results used to evaluate the data model. Final chapter will summarize the overall research and highlights the significance findings of the research.

Chapter 8

Conclusion and Further Work

8.1 Introduction

This chapter provides an overview of the research and how we provide the solution to address the problem of analyzing cancer patient's records which belong to big data category. Furthermore this chapter focuses of limitations and further work of this research.

8.2 Overview of the research

By analyzing patient records we can help the patients to solve their financial issues when they face day by day due to the illness. The key to managing their finances is to fully assess their situation. Keeping track of patient's financial state is important. By using this solution it will be helping patient to understand how much money they need to meet their future expenses. This will help the patient to gain more control over their finances. The financial influence of cancer will effect to different persons in different manner depending on the cancer type, stage and treatment, as well as their financial situation.

When comparing existing solutions given around the world, mostly those are statistical based, data mining is identified as a novel approach to analyze with its ability to analyze big data set dynamically and efficiently. Data mining is identified as best approach to find out hidden patterns within this patient record data set, different data mining techniques have been used to address the problem. To determine the accuracy of the solution different algorithm within the selected techniques are used and compared their efficiency before selecting the best approach to make the conclusions.

8.3 Problem encountered and limitations

In this research we used sample data from APEKSHA Hospital Maharagama and we focused only on breast cancer, thyroid cancer and Leukemia patients. The limited sample size was an obstacle to build the prediction model.

8.4 Further Work

This research has been carried out to help the patients suffering from cancer to plan their financial matters. For this purpose we use the patient records from the hospital. There were several unused variables in this data set such as readmission<30 days ,test results of the lab test, no of admission ect. By using those we can further extend this research such to find the patients readmission with in short period of time or to help the hospital management to find out future high cost patients or predicting the length of stay for a patient or to predicting of treatment for patients base on their cancer severity.

8.5 Summary

This chapter concludes the thesis by describing the solution given with data mining to analyze the cancer patient records and how it can be enhance further to improve the level of accuracy.

References

[1] M. H. Tekieh and B. Raahemi, "Importance of Data Mining in Healthcare: A Survey," 2015, pp. 1057-1062.

[2] M. Durairaj and V. Ranjani, "Data mining applications in healthcare sector a study," *Int. J. Sci. Techno!. Res.,* val. 2, no. 10, pp. 29-35, 2013.

[3] F. Coenen, "Data mining: past, present and future," *Know/. Eng. Rev.,* val. 26, no. 1, pp. 25-29, Mar. 2011.

[4] N. Menachemi and Collum, "Benefits and drawbacks of electronic health record systems," *Risk Manag. Healthc. Policy,* p. 47, May 2011.

[5] "Annual Health Bulletin -Sri Lanka (2014)." Medical Statistics Unit Ministry of Health, Nutrition and Indigenous Medicine- Sri Lanka, 2016.

[6] B. . J. Berkman and S. E. Sampson, "Psychosocial Effects of Cancer Economics on Patients and Their Families." CANCER Supplement November 1, 1993, Volume 72. No. 9, 18-Jun-1993.

[7] L. Sharp and A. Timmons, "The financial impact of a cancer diagnosis." .

[8] L. Sun, C. Liu, C. Guo, H. Xiong, and Y. Xie, "Data-driven Automatic Treatment Regimen Development and Recommendation," 2016, pp. 1865-1874.

[9] R. Ramanayake, D.P. Perera, A. H. W. De Silva, and R. D. N. Sumanasekara, "Patient held medical record: solution to fragmented health care in Sri Lanka," *Thehealth,* val. 4, no. 3, pp. 51-57, 2013.

[10] P. Ahmad, S. Qamar, and S. Q. A. Rizvi, "Techniques of data mining in healthcare: A review," *Int. J. Comput. Appl.,* val. 120, no. 15, 2015.

[11] D. Bertsimas *eta!.,* "Algorithmic Prediction of Health-Care Costs," *Oper. Res.,* val. 56, no. 6, pp. 1382-1392, Dec. 2008.

[12] S. Purdy, "Avoiding hospital admissions," *What Does Res. Evid. Say,* pp. 7-8, 2010.

[13] A. Marquardt *eta!.,* "HealthSCOPE: An Interactive Distributed Data Mining Framework for Scalable Prediction of Healthcare Costs," 2014, pp. 1227-1230.

[14] N. Meadem, N. Verbiest, K. Zolfaghar, J. Agarwal, S.-C. Chin, and S. B. Roy, "Exploring preprocessing techniques for prediction of risk of readmission for congestive heart failure patients," in *Data Mining and Healthcare (DMH), at International Conference on Knowledge Discovery and Data Mining (KDD),* 2013.

[15] S. Cao, J. Li, X. Zhang, Y. Tian, J. Zhao, and W. Xue, "A decision-tree-based analysis of the factors influencing single disease costs," in *Systems and Informatics {ICSA/), 2012 International Conference on,* 2012, pp. 2251-2254.

[16] M.S. Bhuvan, A. Kumar, A. Zafar, and V. Kishore, "Identifying Diabetic Patients with High Risk of Readmission," *ArXiv Prepr. ArXiv160204257,* 2016.

[17] F. S. Gharehchopogh and Z. A. Khalifelu, "Neural network application in diagnosis of patient: a case study," in *Computer Networks and Information Technology {ICCNJT}, 2011 International Conference on,* 2011, pp. 245-249.

[18] A. Alsayat and H. El-Sayed,"Efficient genetic K-Means clustering for health care knowledge discovery," in *Software Engineering Research, Management and Applications (SERA}, 2016 IEEE 14th International Conference on,* 2016, pp. 45-52.

[19] H. Elshazly, A. T. Azar, A. El-Korany, and A. E. Hassanien,"Hybrid system for lymphatic diseases diagnosis," in *Advances in Computing, Communications and Informatics (JCACCJ}, 2013 International Conference on,* 2013, pp. 343-347.

[20] D. J. Tufts-Conrad, A. N. Zincir-Heywood, and D. Zitner,"SOM: feature extraction from patient discharge summaries," in *Proceedings of the 2003 ACM symposium on Applied computing,* 2003, pp. 263-267.

[21] S. T. Moturu, W. G. Johnson, and H. Liu,"Predicting future high-cost patients: A real-world risk modeling application," in *Bioinformatics and Biomedicine, 2007. BIBM 2007. IEEE International Conference on,* 2007, pp. 202-208.

[22] H. Jasti, E. M. Mortensen, D. S. Obrosky, W. N. Kapoor, and M. J. Fine,"Causes and Risk Factors for Rehospitalization of Patients Hospitalized with Community-Acquired Pneumonia," *Clin. Infect. Dis.,* vol. 46, no. 4, pp. 55 556, Feb. 2008.

[23] X. Xiao and S. Chiusano,"Analysis of Medical Treatments Using Data Mining Techniques.," *IEEE Intel/. Inform. Bull.,* vol. 15, no. 1, pp. 30-31, 2014.

[24] M. Herland, T. M. Khoshgoftaar, and R. Wald, "A review of data mining using big data in health informatics," *J. Big Data,* vol. 1, no. 1, p. 1, 2014.

[25] D. Roosan, M. Samore, M. Jones, Y. Livnat, and J. Clutter,"Big-Data Based Decision-Support Systems to Improve Clinicians' Cognition," 2016, pp. 285-288.

[26] K. Zolfaghar, N. Meadem, A. Teredesai, S. B. Roy, S.-C. Chin, and B. Muckian,"Big data solutions for predicting risk-of-readmission for congestive heart failure patients," in *Big Data, 20131EEE International Conference on,* 2013, pp. 64-71.

[27] M. Ojha and K. Mathur,"Proposed application of big data analytics in healthcare at Maharaja Yeshwantrao Hospital," in *2016 3rd MEC International Conference on Big Data and Smart City (JCBDSC},* 2016, pp. 1-7.

[28] G. Vaishali and V. Kalaivani, "Big data analysis for heart disease detection system using map reduce technique," in *Computing Technologies and Intelligent Data Engineering (JCCTIDE}, International Conference on,* 2016, pp. 1-6.

[29] A. Chandiok and D. K. Chaturvedi, "Cognitive Decision Support System for medical diagnosis," in *Computational Techniques in Information and Communication Technologies (JCCTICT), 2016 International Conference on,* 2016, pp. 337-342.

[30] H. Ellouzi, H. Ltifi, and M. B. Ayed, "New Multi-Agent architecture of visual Intelligent Decision Support Systems application in the medical field," in *Computer Systems and Applications {AICCSA), 2015 IEEE/ACS 12th International Conference of,* 2015, pp. 1-8.

[31] H. Wang *eta/.,* "Using the LACE index to predict hospital readmissions in congestive heart failure patients," *BMC Cardiovasc. Disord.,* val. 14, no. 1, p. 1, 2014.

[32] E. W. Lee, "Selecting the Best Prediction Model for Readmission," *J. Prev. Med. Pub. Health,* vol. 45, no. 4, pp. 259-266, Jul. 2012.

[33] J. O. Kang, S.-H. Chung, and Y.-M. Suh, "Prediction of hospital charges for the cancer patients with data mining techniques," *J. Korean Soc. Med. Inform.,* val. 15, no. 1, pp. 13-23, 2009.

[34] Y. Chechulin, A. Nazerian, S. Rais, and K. Malikov, "Predicting patients with high risk of becoming high-cost healthcare users in Ontario (Canada)," *Healthc. Policy,* vol. 9, no. 3, p. 68, 2014.

[35] J. S. Huang, Y. F. Chen, and J. C. Hsu, "Design of a Clinical Decision Support Model for Predicting Pneumonia Readmission," 2014, pp. 1179-1182.

[36] X. Zhou, H. Han, I. Chankai, A. Prestrud, and A. Brooks, "Approaches to text mining for clinical medical records," in *Proceedings of the 2006 ACM symposium on Applied computing,* 2006, pp. 235-239.

[37] H. Champion, N. Pizzi, and R. Krishnamoorthy, "Tactical Clinical Text Mining for Improved Patient Characterization," 2014, pp. 683-690.

[38] B. Ray, "IEEE ICHI Healthcare Data Analytics Challenge," 2015, pp. 523-524.

[39] G. Bruno, T. Cerquitelli, S. Chiusano, and X. Xiao, "A Clustering-Based Approach to Analyse Examinations for Diabetic Patients," 2014, pp. 45-50.

[40] K. Zolfaghar *eta/.,* "Predicting risk-of-readmission for congestive heart failure patients: A multi-layer approach," *ArXiv Prepr. ArXiv13062094,* 2013.

[41] K. Zolfaghar, J. Agarwal, D. Sistla, S.-C. Chin, S. Basu Roy, and N. Verbiest, "Risk-a-meter: an intelligent clinical risk calculator," in *Proceedings of the 19th ACM SJGKDD international conference on Knowledge discovery and data mining,* 2013, pp. 1518-1521.

[42] X. Liu, R. Lu, J. Ma, L. Chen, and B. Qin, "Privacy-Preserving Patient-Centric Clinical Decision Support System on Na·ive Bayesian Classification," *IEEE J. Biomed. Health Inform.,* val. 20, no. 2, pp. 655-668, Mar. 2016.

[43] M. B. Hernandez *et al.,* "Predictors of 30-Day Readmission in Patients Hospitalized With Decompensated Heart Failure: Predicting freedom from heart failure readmission," *Clin. Cardia/.,* val. 36, no. 9, pp. 542-547, Sep. 2013.

[44] D. Zikos, K. Tsiakas, F. Qudah, V. Athitsos, and F. Makedon, "Evaluation of classification methods for the prediction of hospital length of stay using medicare claims data," in *Proceedings of the 7th International Conference on PErvasive Technologies Related to Assistive Environments,* 2014, p. 54.

[45] P. Braga, F. Portela, M. F. Santos, and F. Rua, "Data mining models to predict patient's readmission in intensive care units," in *JCMRT 2014-Proceedings of the 6th International Conference on Agents and Artificial Intelligence,* 2014.

[46] S. Chatterjee, Q. Xie, and K. Dutta, "A predictive modeling engine using neural networks: Diabetes management from sensor and activity data," in *e-Health Networking, Applications and Services (Healthcom), 2012 IEEE 14th International Conference on,* 2012, pp. 230--237.

[47] R. Caruana, Y. Lou, J. Gehrke, P. Koch, M. Sturm, and N. Elhadad, "Intelligible Models for HealthCare: Predicting Pneumonia Risk and Hospital 30-day Readmission," 2015, pp. 1721-1730.

[48] S. S. Disanayaka, H. G. N. De Zoysa, S. Senaratne, and K. R. D. De Silva, "Analysis of medical records of patients with 'other neurological diseases' admitted to Lady Ridgeway Hospital," *Sri Lanka J. Child Health,* val. 37, no. 4, 2009.

[49] H. M. Ratnayake, "Negotiating privacy, confidentiality and security issues pertaining to electronic medical records in Sri Lanka: A comparative legal analysis," *Sri Lanka J. Bio-Med. Inform.,* val. 4, no. 2, Dec. 2013.

[SO] A. M. M. Riyaz, "Study of factors affecting readmission of elders," Post Graduate of Medicine (PGIM), University of Colombo, Sri Lanka, 2008.

[51] B. M. Bai, N. Mangathayaru, and B. P. Rani, "An Approach to Find Missing Values in Medical Datasets," 2015, pp. 1-7.

[52] T. Pang-Ning, M. Steinbach, and V. Kumar, *Introduction to Data Mining.* Addison-Wesley.

[53] "https://www.invensis.net/blog/data-processing/12-data-mining-tools-techniques."

[54] "http://www.cs.waikato.ac.nz/ml/weka/.".

[55] "http://www.statisticallysignificantconsulting.com/RegressionAnalysis.htm/." .

[56] "http://wiki.pentaho.com/display/DATAMINING/SMOreg/.".

[57] "http://math.tutorvista.com/statistics/correlation.html/." .

[58] "http://www.theanalysisfactor.com/assessing-the-fit-of-regression-models/." .

[59] "http://www.studytonight.com/java/features-of-java.php/." .

Appendix- A

Variable available in the data set collected from Apeksha Hospital Maharagama.

Rec.ID

GenderAge

Diagnosis

Stage

patient state

treatment

no of lab procedures

DOSAGE

FREQUENCY

Surgery

blood transfusion

Conclusion

radio_theropy

Tamoxifen

cytarabine

Cladribine

Methotrexate

Herceptin

SF4

Vincristine(VCR)

Fluorouracil (5FU)

Imatinib 400mg

Doxorubicin

docetaxel

Paclitaxel

Anastrozole

Asparaginase

Dexakepyon

Dasatinib

BMP

radio Iodine

thyroxine

immunization

Cyclophosphamide

Appendix- B

Figure 6.4.1 - Applying Numeric to Norninal filter before attribute selection.

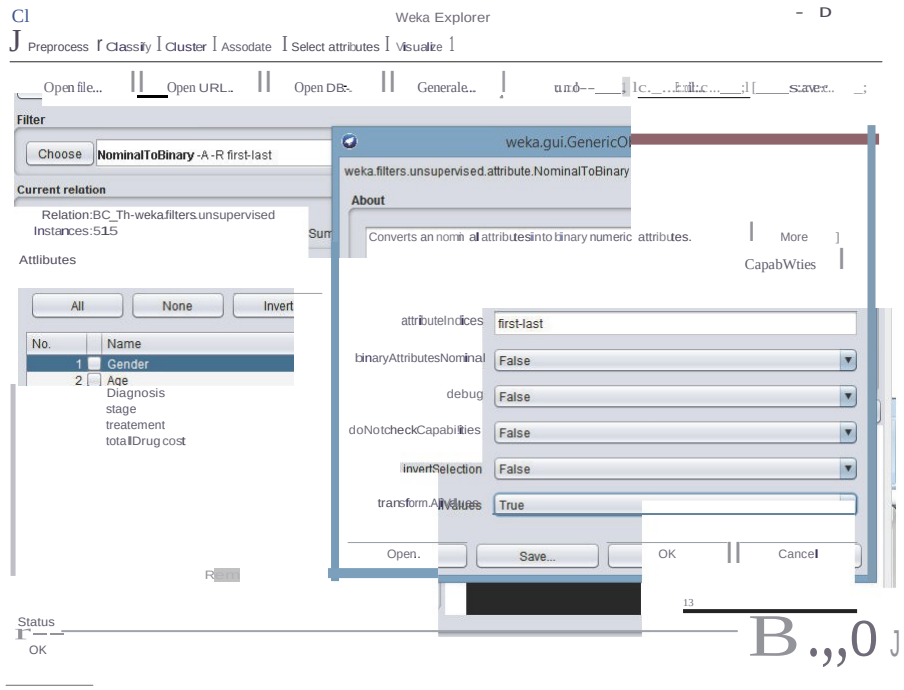

Figure 6.42 Apply NorninalToBinary Filter

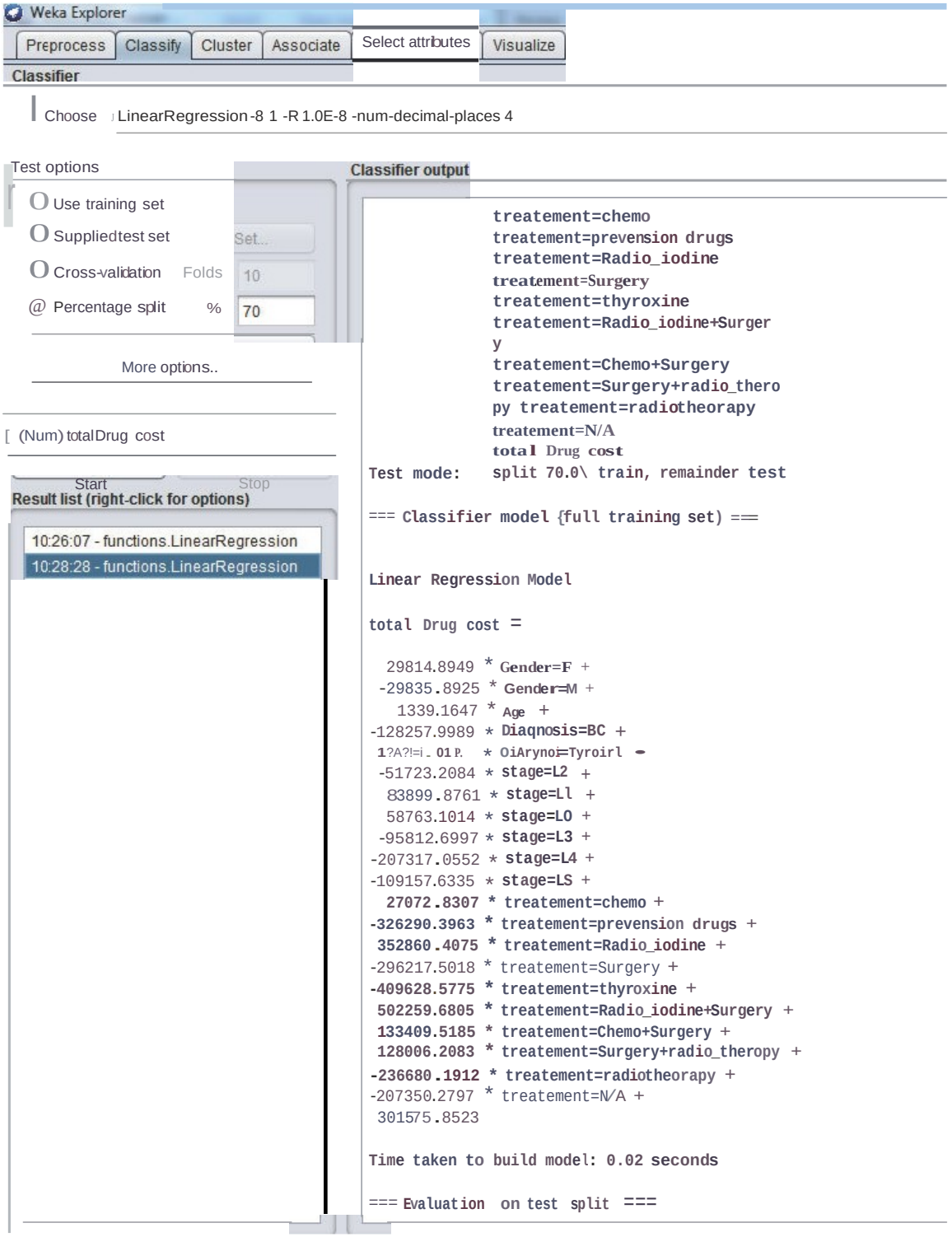

Figure 6.3 -After applying Linear regression model

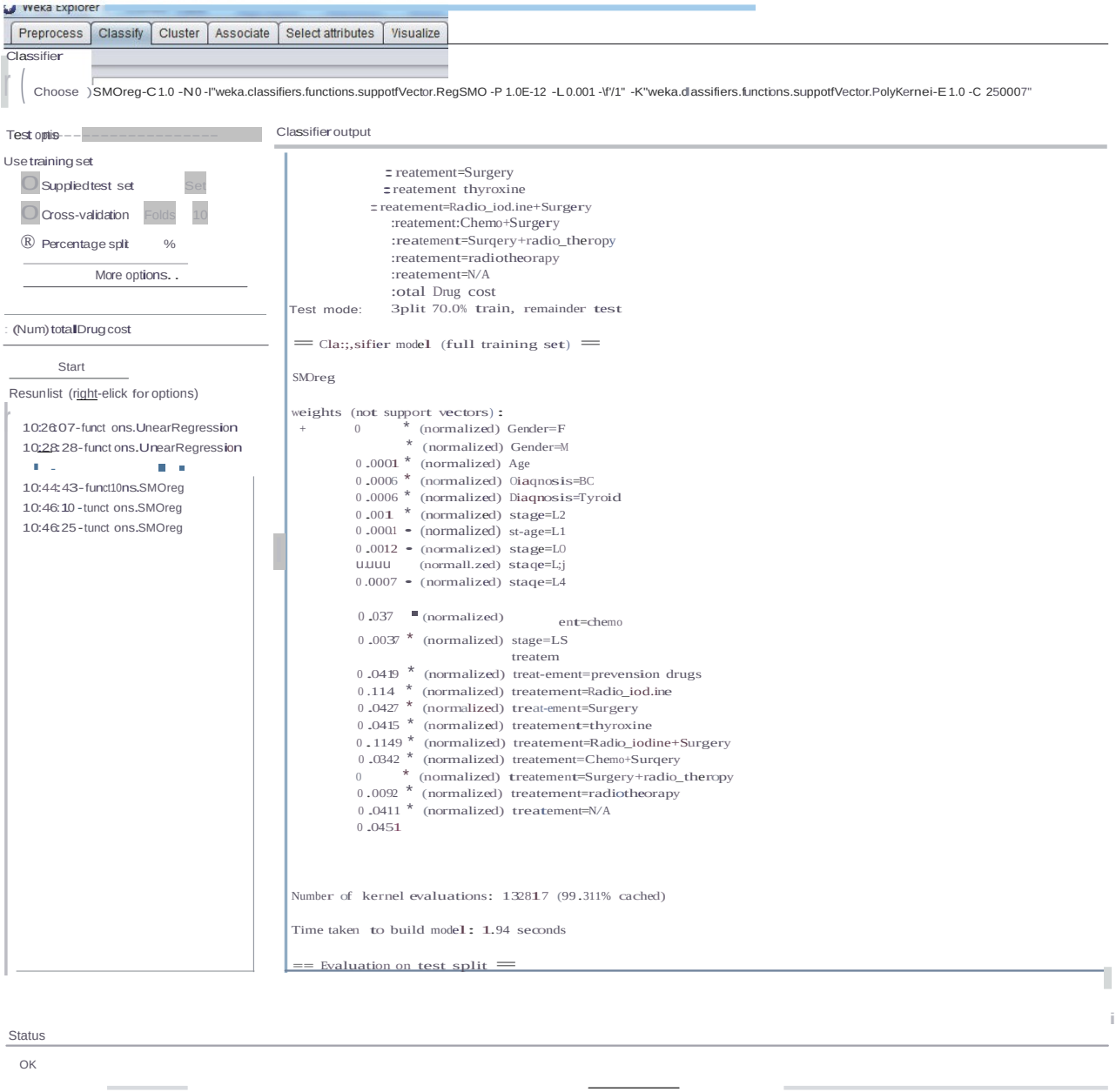

Figure 6.4 - After applying SMO regression model

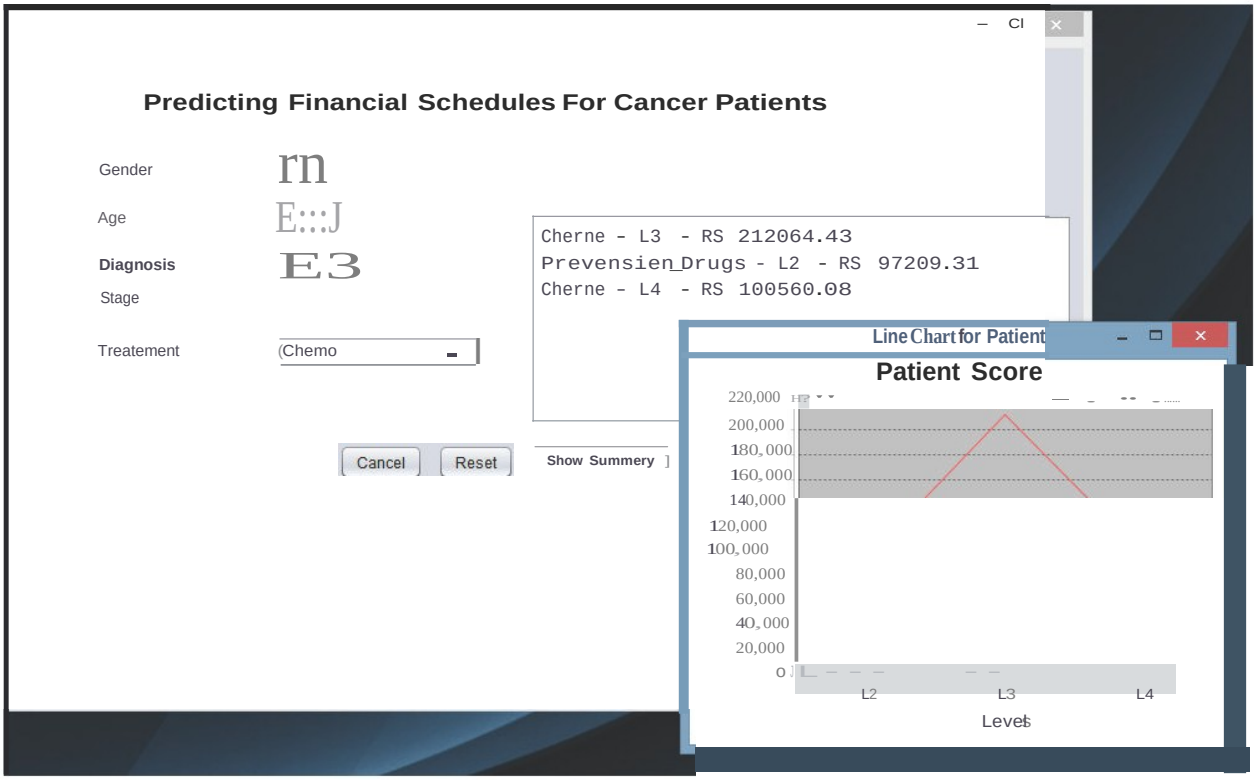

Figure 6.6-Graphical view of the application

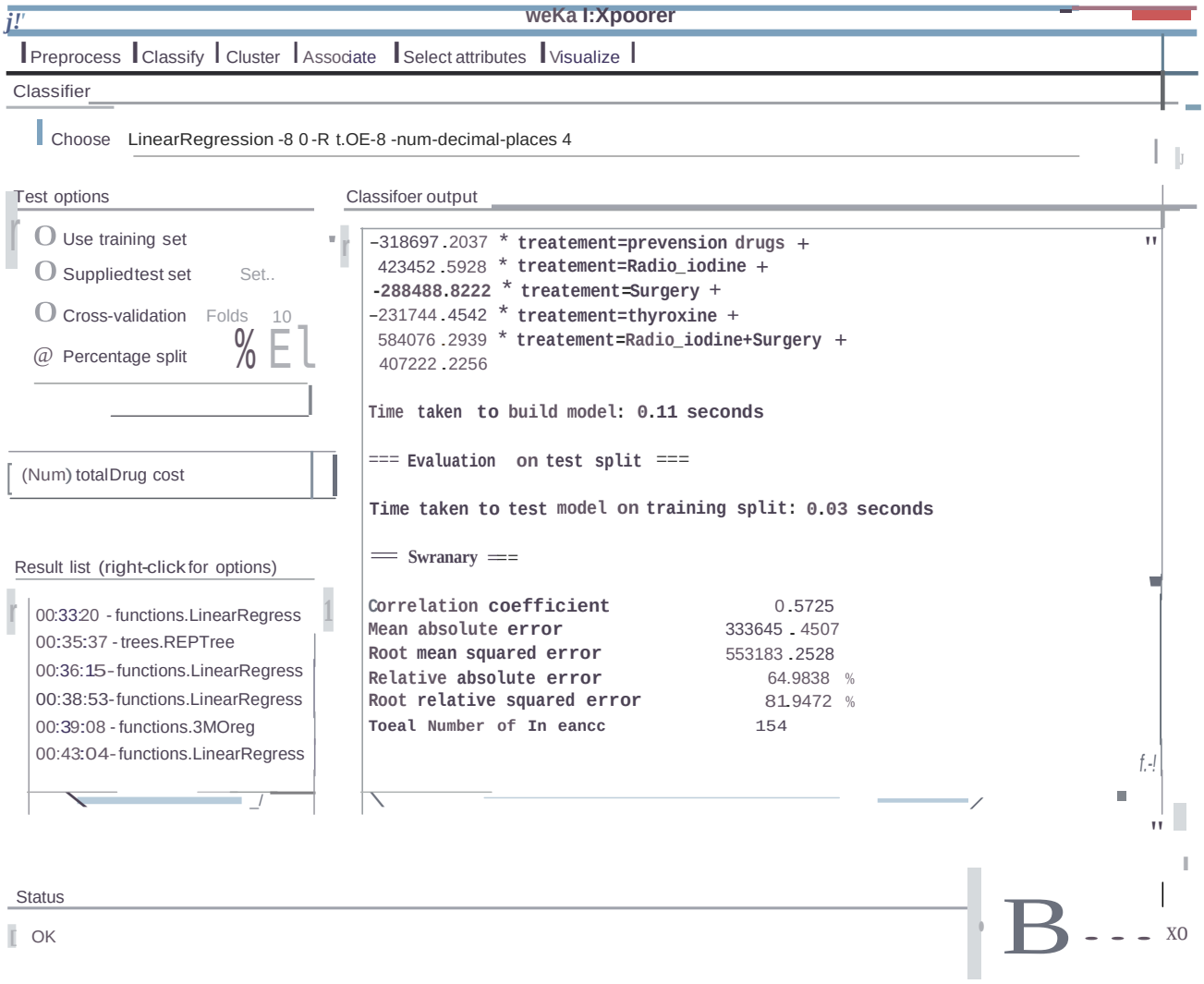

Figure 7.1- Evaluation results of the multiple linear regression model

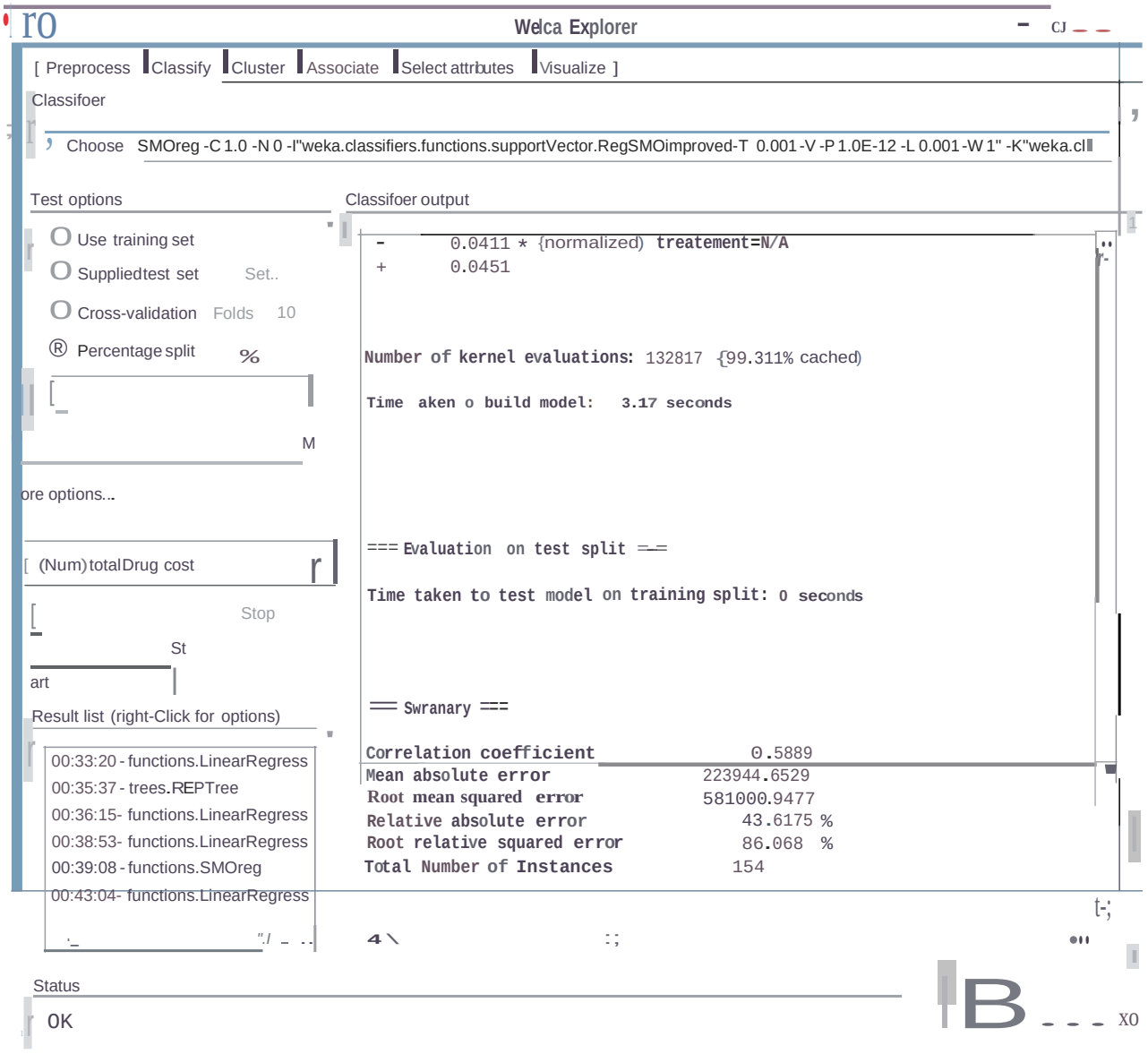

Figure 7.2- Evaluation results of the SMO regression model

Appendix- C

Code for Building the Linear Regression Model

```
package com.cancer;

import weka.core.Instances;

import weka.core.converters.ConverterUtils.DataSource;

import weka.classifiers.functions.LinearRegression;

/!import weka.classifiers.functions.SMOreg;

public class Regression{

        public static void main(String args[]) throws Exception{

                //load dataset

                DataSource source= new DataSource("E:/Msc/project_on_28-
4/patient_data_reduce.arff");

                Instances dataset= source.getDataSet();

                //set class index to the last attribute

                dataset.setClasslndex(dataset.numAttributes()-1);

                //build model

                LinearRegression lr =new LinearRegression();

                lr.buildClassifier(dataset);

                //output model

                System.out.println(lr);

        }}
```

Code for Predicting the Cost For Future Medicine

```
package com.cancer; import

java.io.FileWriter; import

java.io.IOException;

import java.text.DecimalFormat;

import java.util.ArrayList;

import java.util.Arrays;

import java.util.Collections;

import java.util.List;

import weka.classifiers.functions.LinearRegression;

import weka.core.Instance;

import weka.core.Instances;

import weka.core.converters.ConverterUtils;

public class WekaService {

   private static String FILE_LOCATION = "E:I";

   private static String FILE_EXTENTION = ".arff";

   private static String MODEL_PATH = "E:/final presentation/Cancer/datamodelll-
5/bc_th_model_test.model";

    SysCache cache = SysCache.getlnstance();

   private static DecimalFormat df2 = new DecimalFormat(".##");

   public List<String> writeToArffFormat(String gender, int age, String diagnosis,

         String stage, String treatment) throws IOException, Exception {

      String modStage = "";
```

```
String preTreatMent= treatment;
String aftTreatMent =treatment;
List<String> stageList =new ArrayList<String>();
stageList.add("LO");
stageList.add("Ll");
stageList.add("L2");
stageList.add("L3");
stageList.add("L4");
stageList.add("L5");
stageList.add("L6");
cache.clearPredictedValues();
List<String> statusList =new ArrayList<String>();
statusList.add(generateWithContent(gender, age, diagnosis, stage, treatment,
FileType.CURRENT));
if (diagnosis.equals(Diagnosis.Tyroid.toString())) {
    switch (stage) {
        case "LO":
            aftTreatMent = Treatment.THYROXINE.toString();
            break;
        case "Ll":
            preTreatMent=  Treatment.THYROXINE.toString();
            aftTreatMent  =  Treatment.THYROXINE.toString();
            break;
```

```
                case "L2":
                   if (treatment.equals(Treatment.RADIO_IODINE.toString())) {
                      preTreatMent= Treatment.RADIO_IODINE.toString();
                      aftTreatMent  = Treatment.RADIO_IODINE.toString();
                   } else if (treatment.equals(Treatment.CHEMO_SURGERY.toString())
                         || treatment.equals(Treatment.SURGERY.toString())) {
                      preTreatMent= Treatment.PREVENSION_DRUGS.toString();
                      aftTreatMent  = Treatment.CHEMO.toString();
                   } else if (treatment.equals(Treatment.RADION_THEROPY.toString())) {
                      preTreatMent= Treatment.PREVENSION_DRUGS.toString();
                      aftTreatMent  = Treatment.SURGERY.toString();
                   } else if
(treatment.equals(Treatment.SURGERY_RADIO_THEROPY.toString())) {
                      preTreatMent= Treatment.PREVENSION_DRUGS.toString();
                      aftTreatMent  = Treatment.RADION_THEROPY.toString();
                   }
                   break;
                case "L3":
                   preTreatMent= Treatment.PREVENSION_DRUGS.toString();
                   aftTreatMent  = Treatment.CHEMO.toString();
                   break;
                case "L4":
                   preTreatMent= Treatment.PREVENSION_DRUGS.toString();
```

```
            aftTreatMent  = Treatment.CHEMO.toString();
            break;
        case "L5":
            preTreatMent= Treatment.PREVENSION_DRUGS.toString();
            aftTreatMent  = Treatment.CHEMO.toString();
            break;
    }
} else if (diagnosis.equals(Diagnosis.BC.toString())) {
    switch (stage) {
        case "LO":
            aftTreatMent  = Treatment.PREVENSION_DRUGS.toString();
            break;
        case "Ll":
            preTreatMent= Treatment.PREVENSION_DRUGS.toString();
            aftTreatMent  = Treatment.CHEMO.toString();
            break;
        case "L2":
            if (treatment.equals(Treatment.CHEMO.toString())) {
                preTreatMent= Treatment.PREVENSION_DRUGS.toString();
                aftTreatMent  = Treatment.SURGERY.toString();
            } else if (treatment.equals(Treatment.CHEMO_SURGERY.toString())
                    || treatment.equals(Treatment.SURGERY.toString())) {
                preTreatMent= Treatment.PREVENSION_DRUGS.toString();
```

```
                aftTreatMent  = Treatment.CHEMO.toString();
              } else if (treatment.equals(Treatment.RADION_THEROPY.toString())) {
                preTreatMent= Treatment.PREVENSION_DRUGS.toString();
                aftTreatMent  = Treatment.SURGERY.toString();
              } else if
(treatment.equals(Treatment.SURGERY_RADIO_THEROPY.toString())) {
                preTreatMent= Treatment.PREVENSION_DRUGS.toString();
                aftTreatMent  = Treatment.RADION_THEROPY.toString();
              }
             break;
           case "L3":
             preTreatMent= Treatment.PREVENSION_DRUGS.toString();
             aftTreatMent  = Treatment.CHEMO.toString();
             break;
           case "L4":
             preTreatMent= Treatment.PREVENSION_DRUGS.toString();
             aftTreatMent  = Treatment.CHEMO.toString();
             break;
           case "L5":
             preTreatMent= Treatment.PREVENSION_DRUGS.toString();
             aftTreatMent  = Treatment.CHEMO.toString();
             break;
         }
```

```
        }
        //previous
        modStage = getPrevious(stage, stageList); System.out.println("Pre-" +
        modStage); statusList.add(generateWithContent(gender, age, diagnosis,
        modStage,
preTreatMent, FileType.PREVIOUS));


        //after
        modStage = getNext(stage, stageList);
        System.out.println("After-" + modStage);
        statusList.add(generateWithContent(gender, age, diagnosis, modStage,
aftTreatMent, FileType.AFTER));
        statusList.removeAll(Collections.singleton(null));
        return statusList;
    }
   public String generateWithContent(String gender, int age, String diagnosis,
        String stage, String treatment, FileType fileType) throws IOException,
Exception {
        String fileLocation = FILE_LOCATION;
        String predictValue =null;

        if (fileType == FileType.PREVIOUS) {
            fileLocation += "predict_file_previous";
        } else if (fileType == FileType.CURRENT) {
```

```
    fileLocation += "predict_file";
} else if (fileType == FileType.AFTER) {
    fileLocation += "predict_file_after";
}
fileLocation += FILE_EXTENTION;
FileWriter writer= new FileWriter(fileLocation);
List<CancerEntry> cancerEntrys = Arrays.asList(
        new CancerEntry(gender, age, diagnosis, stage, treatment, "?")
);

/!if stage is empty that means previous of after stages are not available
if (stage != "") {
    //for header
    ArffUtils.writeHeaders(writer);
    for (CancerEntry d : cancerEntrys) {
        List<String> list= new ArrayList<>();
        list.add(d.getGender());
        list.add(String.valueOf(d.getAge()));
        list.add(d.getDiagnosis());
        list.add(d.getStage());
        list.add(d.getTreatment());
        list.add(d.getCost());
```

```
            AtffUtils.writeLine(writer, list);
        }
    } else {
        AtffUtils.writeLine(writer, Arrays.asList(" "));
    }

    writer.flush();
    writer.close();

    //Only stage is available we need to check the predicts values
    // stage is empty then file is empty
    if (stage != "") {
        //Predict Values for given Identifier
        predictValue = predictIdentifier(fileLocation, fileType, stage, treatment);
    }
    return predictValue;
}

public String predictIdentifier(String filePath, FileType fileType, String stage,
String treatment) throws Exception {
    //load model
LinearRegression lr2 = (LinearRegression)
weka.core.SerializationHelper.read(MODEL_PATH);
String predictValue =null;
```

```
//load new dataset
ConverterUtils.DataSource sourcel =new ConverterUtils.DataSource(filePath);

Instances testDataset = sourcel.getDataSet();

//set class index to the last attribute
testDataset.setClassindex(testDataset.numAttributes() - 1);
//loop through the new dataset and make predictions
    for (inti = 0; i < testDataset.numinstances(); i++) {
    //get class double value for current instance
    double actualValue = testDataset.instance(i).classValue();
    //get Instance object of current instance
    Instance newinst = testDataset.instance(i);
    //call classifyInstance, which returns a double value for the class
    double predLR = lr2.classifyinstance(newinst);
    predLR = Math.abs(Double.parseDouble(df2.format(predLR)));
    System.out.println(actualValue + "\t" + predLR);
    //predictValue = fileType.toString() + II- II+ actualValue +II , II+ predLR;
    predictValue =treatment+ ‖ - ‖ + stage+ ‖ - RS ‖ + Double.toString(predLR);
    cache.setPredictedValue(stage, predLR);
  }
  return predictValue;
}
```

```
    public String getNext(String uid, list<String> stageList) {
        int idx = stageList.indexOf(uid);
        if (idx < 0 ‖ idx + 1 == stageList.size()) {
            return"";
        }
        return stageList.get(idx + 1);
    }

    public String getPrevious(String uid, list<String> stageList) {
        int idx = stageList.indexOf(uid);
        if (idx <= 0) {
            return"";
        }
        return stageList.get(idx - 1);
    }

}
```