UNIVERSITY OF NAIROBI

School of Computing and Informatics

A Data Mining approach to private healthcare services demand forecast in Nairobi County

By

IRERI ANDREW MURIITHI

Supervisor

Dr. Lawrence Muchemi

A research Project Report submitted in partial fulfilment of the requirements of the Masters of Science degree in Computer science of University of Nairobi, Kenya.

November 2014

DECLARATION

Student

I hereby declare that this project report is my own original work and has, to the best of my knowledge, not been submitted to any other institution of higher learning.

Sign: …………………………………….   Date: ……………………………………

Ireri Andrew Muriithi

P58/76112/2012

Supervisor

This project report has been submitted as a partial fulfilment of requirements for the Masters of Science in Computer Science of the University of Nairobi with my approval as the university supervisor.

Sign: …………………………………….   Date: ……………………………………

Dr. Lawrence Muchemi

SCHOOL OF COMPUTING AND INFORMATICS

ABSTRACT

One of the most important indices of defining general welfare and quality-of-life of people in the world is physical and mental health of individuals. Health care facilities in Kenya have always aimed to provide health care for all residents using a fair access policy that is characterized by providing the right service at the right time in the right place. Even though due to computerization, healthcare industry in Kenya today generates large amounts of complex data about patients, hospitals resources, disease diagnosis, electronic patient records, and medical devices, there does not exist an intelligent system that can mine this big data and provide analytical patterns on healthcare services demand forecast in Kenya. Health care managers and planners therefore must make future decisions about healthcare services delivery without knowing what will happen in the future. Forecasts would enable the managers to anticipate the future demand and plan accordingly.

This study aimed at examining Data Mining as an approach to private health services demand forecast in Nairobi County in Kenya. Data mining was used as it brings in a set of tools and techniques that can be applied to the big data to discover hidden patterns that provide healthcare professionals an additional source of services demand forecast knowledge for making decisions. A supervised Artificial Neural Network-based model for private health care services demand forecast was developed. A prototype was then developed from the model and its performance evaluated by applying the actual private services demand data from the DHIS2 Kenya system to predict the demand for 3 years in advance. The model was trained under the WEKA environment and predicted the demand for health services in various categories for private health care providers in Nairobi county Kenya. The test results showed that the forecast year-by-year approach was more suitable and efficient for years ahead demand forecasting. Forecast results demonstrated that the model performed remarkably well with increased number of actual data and iterations. Artificial Neural Networks model gave a more accurate forecast results with 4% Mean Percentage Error as compared to alternative methods of demand forecasting whose error was above 6%.

The established private health care demand forecast model gives the health service providers optimal decisions they can make today about the private healthcare business tomorrow.

Key Words: Data Mining, Artificial Neural Network, ANN, Forecast, Private Health Care, Decision Making, WEKA, Linear Regression, Time Series

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to the following:

- The Almighty God for His mercies and grace, a gift of life and good health to pull through.
- The management of University of Nairobi, which gave me the environment and learning resources to undertake my Master's degree and write this proposal
- Dr. Lawrence Muchemi, who, being my supervisor in this work, supported me all though and lead me through the academic path.
- My friends and colleagues at work Mercy Nyakio, Judy Nganga, Kezia Kinyajui, Regina Lasiti, Solomon Kamau, Esther Ndungu and Reagan Muriuki who gave me unfailing moral support to pull through this program

Table of Contents

List of Tables

List of Figures

List of Abbreviations

API – Application Programming Interface

CRISP-DM- Cross Industry Standard Process for Data Mining

DM- Data mining

ICT- Information Communication Technology

GIS - Geographical Information Systems

KDD - Knowledge Discovery in Databases

OP – Out Patient

SEMMA- Sample Explore Modify Model and Assess

XML - Extensible Markup Language

CHAPTER 1: INTRODUCTION

1.1 BACKGROUND OF STUDY

One of the most important indices of defining general welfare and quality-of-life of people in the world is physical and mental health of individuals. Health care facilities at any region in Kenya can be divided into two main categories that are known as public and private healthcare facilities. The public healthcare facilities are funded and managed by the government of Kenya while private facilities are privately funded and managed. Health authorities have always aimed to provide health care for all residents using a fair access policy that is characterized by providing the right service at the right time in the right place. There are several issues (challenges) that make such health policy difficult to be implemented. One of these issues is related to the relationship between distance to health services and the need for health care. The use of health services is influenced also by other factors including financial status, time constrains, social inconveniences and the psychological stress of journey to the health services (Abdulkader, 2006)

Every day, health care managers must make decisions about service delivery without knowing what will happen in the future. Forecasts would enable them to anticipate the future and plan accordingly. Demand forecasting is a key business functions for retailers, manufacturers and service providers. Good forecasts are the basis for short-, medium-, and long-term planning, and are essential input to all types of service production systems. Forecasts have two primary uses: to help managers plan the system, and also to help them plan the use of the system. Planning the system itself is long-range planning: about the kinds of services supplied and the number of each to offer, what facilities and equipment to have, which location optimizes service delivery to the particular patient population, and so on (Yasar, 2005).

Lack of accurate and credible information about the demand for essential health products costs lives. Gaps and weaknesses in demand forecasting result in a mismatch between supply and demand – which in turn leads to both unnecessarily high prices and supply shortages. Although it is only one step in the long and often complicated supply chain, demand forecasting represents a key point of decision-making for both buyers and suppliers. If demand forecasting isn’t done well

– given inherent uncertainties, particularly for newer markets – the rest of the supply chain cannot be efficiently mobilized to deliver treatment (Neelam, 2006).

Healthcare industry today generates large amounts of complex data about patients, hospitals resources, disease diagnosis, electronic patient records, medical devices etc. The large amounts of data is a key resource to be processed and analyzed for knowledge extraction that enables support for cost-savings and decision making. Data mining brings a set of tools and techniques that can be applied to this processed data to discover hidden patterns that provide healthcare professionals an additional source of knowledge for making decisions as far as demand forecasting is concerned (Prasanna, Kuo-Wei, & Jaideep, 2011).

Data mining is an automated approach for discovering or inferring hidden patterns or knowledge buried in data. 'Hidden' means patterns that are not made apparent through casual observation. Data Mining is an interdisciplinary field that combines artificial intelligence, computer science, machine learning, database management, data visualization, mathematic algorithms, and statistics. This technology provides different methodologies for decision making, problem solving, analysis, planning, diagnosis, detection, integration, prevention, learning, innovation, forecasting, and estimation (Salim & etal, 2013). Data mining and analysis has become increasingly essential as financial pressures have heightened the need for healthcare organizations to make decisions based on the analysis of clinical and financial data.

Much of causing factors of the diseases are essentially spatial; i.e. their distribution and concentration vary in different locations. Geo-spatial data mining, a subfield of data mining, is a process to discover interesting and potentially useful spatial patterns embedded in spatial databases. Efficient GIS tools to extract information from massive geo-spatial datasets are crucial for organizations to own, generate, and manage geo-spatial datasets (Abdulkader, 2006).

The presented research proposal proposes a new sophisticated healthcare services demand forecasting solution based on data mining techniques for private healthcare service providers in Nairobi County.

1.2 STATEMENT OF THE PROBLEM

In Kenya today, health information and healthcare data analytics have been extensively used to measure health indicators, comparative analysis for planning and administration of quality health services and scientific research. However, the current situation is that healthcare managers and planners must make future decisions about healthcare services delivery without knowing what will happen in the future.

Even though healthcare industry in Kenya today generates large amounts of complex data about patients, hospitals resources, disease diagnosis, electronic patient records, and medical devices, there does not exist an intelligent and sophisticated system that can mine this big data and provide analytical patterns on healthcare services demand forecast in Kenya. Forecasts would enable the managers to anticipate the future demand and plan accordingly. Lack of accurate and credible information about the future demand for essential healthcare services can actually costs lives. This is because gaps and weaknesses in demand forecasting result in a mismatch between supply and demand – which in turn leads to both unnecessarily high prices and supply shortages. This situation is bound to get worse if nothing is done presently as the population in Kenya continues to grow at a very high rate.

This project proposes the use of data mining approach to build a healthcare services demand forecasting model in Nairobi County as a solution to this problem. This demand forecasting model feeds into specific decision variables, giving the health service provider an optimal decision he can make today about the private healthcare business tomorrow.

1.3 RESEARCH OBJECTIVES

Overall Objective

To examine data mining as an approach to private health services demand forecast in Nairobi County.

Specific Objectives;

1. To explore the various forecasting methods in healthcare services demand
2. To identify the most suitable healthcare services demand forecasting methods for Nairobi County.

3. To develop and validate a data mining model for forecast of private healthcare services demand for Nairobi County.

1.4 JUSTIFICATION AND SIGNIFICANCE OF THE STUDY

Health care providers can no longer afford to indulge in the ***"build it and they will come"*** fallacy. Though many hospitals and medical centers have operated under Reilly's law of retail gravity - more square footage equals a larger trade area to draw from - they have begun to realize that to be competitive they need to be located conveniently to their customer base. How consumers access the services of managed health care providers is controlled by geographic location.

The majority of retailers still conduct demand forecasting and planning using outdated, often homegrown systems that lack forecasting algorithms and analytical tools. Without using data mining techniques and models, not only is it difficult to design system with better demand forecasts but it is also not efficient to handle large data set with many relations and attributes.

The main goal of this study is to develop a demand forecasting model for private healthcare services in Nairobi County by heavily employing data mining techniques. This model can assist the healthcare managers in the private sector in evidence-based decision making in planning healthcare services to meet the escalating demand.

1.5 SCOPE AND LIMITATION OF THE STUDY

This project will focus on health services demand for Nairobi County as the case study and will not include other counties. Data collected and analyzed will be from this county only and will only be from the private healthcare facilities. The model to be developed will be targeting these private facilities as well. The only healthcare services that will form the basis of forecasts are;

- Outpatient total cases
- Outpatient specific medical cases
- Surgical cases (Cardiac, Cancer, Orthopedic)

CHAPTER 2: LITERATURE REVIEW

2.1 Introduction

Millions of databases have been used in business management, government administration, scientific and engineering data management, and many other applications. This explosive growth in data and databases has generated an urgent need for new techniques and tools that can intelligently and automatically transform the processed data into useful information and knowledge, which provide enterprises with a competitive advantage, working asset that delivers new revenue, and to enable them to better service and retain their customers (Mouhib & Wael, 2008).

Accurately projecting the demand for specific goods and services helps companies to order raw materials and schedule production of those products and services in a timely manner, making it possible to meet consumer needs quickly and efficiently without the need to build up a large inventory that adds to the tax burden of the business (Patrizia & Karin, 2013).

2.2 Demand forecasting in healthcare

According to (Chendroyaperumal, 2009), Capital is the seed to sprout a firm and hence it is the strategic variable at the inception. The firm survives, takes root and grows only when its product-fruits are sought and bought. Thus demand for the product becomes the strategically critical variable after the inception. Therefore all firms strive to forecast the demand for their products or services. Economists, through their varied definitions, generally mean demand (of whatever type) to be the need for the commodity or service (let us call this as a 'product') plus the willingness to buy plus the capacity to buy or the purchasing power or their money income. That is when we refer to the concept 'demand' we are, in fact, referring to a group of three variables namely;

- the need,
- the willingness and
- The money income.

The classic economics definition of ‘demand’ is fairly straightforward – it states that from the perspective of sovereign consumers, demand defines the quantity of a given product they are willing to buy at a given price. Demand forecasting is the process of planning and determining which products will be purchased, where, when, and in what quantities.

However, the term “demand forecasting” as used by various actors in the health supply chain does not often conform to this definition but reflects instead the particular forecasting needs of the player;

- Within international agencies, it is often used to mean “needs forecasting” – e.g. the number of people affected by a disease based on epidemiological data and the proportion of those requiring treatment.
- Funders use it to mean “resource forecasting” to project needs for future financing, usually from the donor community.
- For country programs and buyers, it can range from describing short term budget needs to achieving ambitious government targets.
- In global health programs, it is often used synonymously with “demand creation” to really mean *generating* demand for products that can be used to address public health challenges.
- Finally, suppliers use the term demand forecasting to determine resource requirements in planning for new products and services, or in the case of existing products, to guide their investments in production capacity and raw materials (Neelam, 2006).

Demand forecasting is the area of predictive analytics dedicated to understanding consumer demand for goods or services. That understanding is harnessed and used to forecast consumer demand. To be able to meet consumers’ needs, appropriate forecasting models are vital. Although no forecasting model is flawless, unnecessary costs stemming from too much or too little supply can often be avoided using data mining methods. Using these techniques, a business is better prepared to meet the actual demands of its customers (Balaji, 2011).

A couple of points are worth noting: First, forecasts are not plans or targets. Plans tell us how the future should look or how we would want it to look, while targets are goals used to motivate performance. Forecasts tell us how the future *will most likely* look based on the best data and estimates available. Second, uncertainty is inherent in forecasting. Though more quantitative data definitely improves the quality of forecasts, the costs of additional data collection must be weighed against their benefits (Neelam, 2006).

Forecasts of a group of items (aggregate forecasts) tend to be more accurate than those for individual items. For example, forecasts made for a whole hospital would tend to be more accurate than a departmental forecast, because forecasting errors among a group tend to cancel each other. Finally, it is generally accepted that forecast accuracy decreases as the time horizon (the period covered) increases. Short-range forecasts face fewer uncertainties than longer-range forecasts do, so they tend to be more accurate (Neelam, 2006).

Why Are We Forecasting?

According to (Neelam, 2006), The benefits of accurate forecasting are evident: improved customer service, greater market efficiency resulting from better production planning and lower inventory, adequate supply to customers, and early recognition and supply of future customer needs. The risks of poor forecasting are equally evident: an inefficient market with higher prices and wasted capacity, supply shortages of current drugs, insufficient development of drugs for future needs, and inadequate investment in manufacturing capacity.

Steps in the Forecasting Process

Many forecasting methods are available to health care managers for planning, to estimate future demand or any other issues at hand. However, for any type of forecast to bring about later success, it must follow a step-by-step process comprising five major steps (Yasar, 2005):

1. Goal of the forecast and the identification of resources for conducting it;
2. Time horizon;
3. Selection of a forecasting technique;
4. Conducting and completing the forecast; and
5. Monitoring the accuracy of the forecast.

Identify the Goal of the Forecast

This indicates the urgency with which the forecast is needed and identifies the amount of resources that can be justified and the level of accuracy necessary.

Establish a Time Horizon

Decide on the period to be covered by the forecast, keeping in mind that accuracy decreases as the time horizon increases.

Select a Forecasting Technique

The selection of a forecasting model will depend on the computer and financial resources available in an organization, as well as on the complexity of the problem under investigation.

Conduct the Forecast

Use the appropriate data, and make appropriate assumptions with the best possible forecasting model. Health care managers often have to make assumptions based on experience with a given situation, and sometimes by trial and error. In forecasting, analyzing appropriate data refers to;

a) The availability of relevant historical data;

b) Recognizing the variability in a given data set.

Monitor Accuracy

Since there is an arsenal of techniques available, appropriate for different situations and data representations, healthcare managers must examine their data and circumstances carefully to select the appropriate forecasting approach. Be prepared to use another technique if the one in use is not providing acceptable results. Health care managers must also be alert to how frequently the forecast should be updated, especially when trends or data change dramatically.

2.3 Factors affecting demand of private health care services

In Kenya today demand for private health care services has been steadily going high in the past couple of decades. According to studies done by (Hursh & etal., 1994), the following are the key demand factors affecting private health services.

Income

Higher income families tend to have higher actual use of health services because they are able to afford the cost. But since they can also afford preventive care, they are able to reduce their real need for health services. This is the so called double effect of income.

Household income is the greatest constraint to the use of their services. And, indeed, where there is evidence of household income, great differences are found in private service use by different economic groups: The poor are most likely to use drug sellers, small individual providers, traditional providers, and mission and mosque facilities; while the more affluent classes are more likely to use larger urban private facilities. Income greatly increases the demand of modern private health care facilities than public utilities.

Education

Greater amount of education may enable a person to recognize early symptoms of illness, resulting in the patient's greater willingness to seek early treatment. More educated people are therefore more likely to seek health care and higher-quality care. The higher the level of education, the greater the use of private services. The effect of education on overall health service use is particularly dramatic for mothers. More educated mothers are more likely than others to seek modern antenatal care and care for children's respiratory infections, fever, and diarrhea, as well as immunizations.

Population

The incidence of illness increases with population, and so does the need for health care. The higher the population the more the incidences of illness and the more the numbers that will ultimately seek private healthcare.

Urban areas are more populated and are able to support full-time private for-profit services because of economies of scale, generally higher incomes, greater formal sector employment, and better access to transportation.

2.4 Data mining and demand forecasting in healthcare

Nowadays there is huge amount of data stored in real-world databases and this amount continues to grow fast as it creates both an opportunity and a need for semi-automatic methods that discover the hidden knowledge in such database. If such knowledge discovery activity is successful, discovered knowledge can be used to improve the decision making process of an organization (Salim & etal, 2013).

Data Mining is an interdisciplinary field that combines artificial intelligence, computer science, machine learning, database management, data visualization, mathematic algorithms, and statistics.

It is a term that describes different techniques used in a domain of machine learning, statistical analysis, modeling techniques and data base technologies that can be used in different industries. With a combination of these techniques, it is possible to find different kinds of structures and relations in the data, as well as to derive rules and models that enable prediction and decision making in new situations. It is possible to perform classification, estimation, forecasts, affinity grouping, clustering and description and visualization (Salim & etal, 2013).

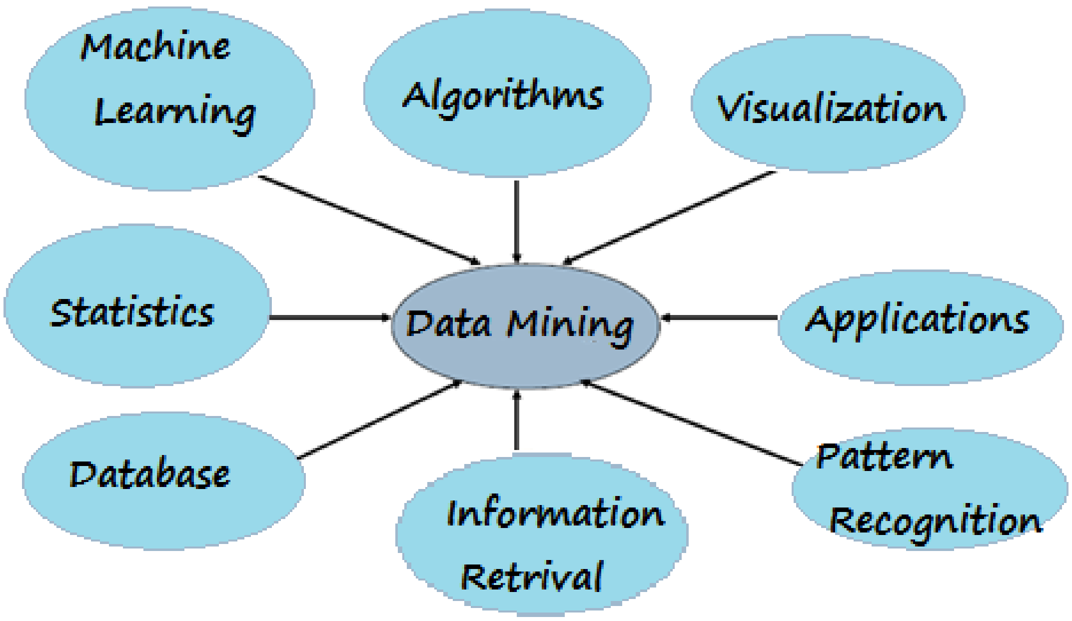


Figure 2.1: Data Mining Architecture

Insights gained from data mining can influence cost, revenue and operating efficiency while maintaining a high level of care. Healthcare organizations that perform data mining are better positioned to meet their long term needs; data can be a great asset to healthcare organizations, but they have to be first transformed into information. Data mining applications also can benefit healthcare providers such as hospitals, clinics, physicians, and patients by forecasting healthcare services demand and hence plan accordingly for expansion programs (Salim & etal, 2013).

Data mining process

A data mining task includes pre-processing, the actual data mining process and post-processing. During the pre-processing stage, the data mining problem and all sources of data are identified, and a subset of data is generated from the accumulated data. To ensure quality the data set is processed to remove noise, handle missing information and transformed it to an appropriate format. A data mining technique or a combination of techniques appropriate for the type of knowledge to be discovered is applied to the derived data set. The last stage is post-processing in which the discovered knowledge is evaluated and interpreted (Mouhib & Wael, 2008).

Knowledge Discovery

Knowledge discovery is a non-tedious procedure for identifying effective and potential benefits amid data. It is known from Fig. 2 below that data mining is one of the important processes of knowledge discovery. From the definitions by the scholars, it is clear that the usage of data mining is an analysis process within a series of knowledge discovery. As time changes the term ‘‘data mining’’ gradually replaces ‘‘knowledge discovery’’ (Salim & etal, 2013).

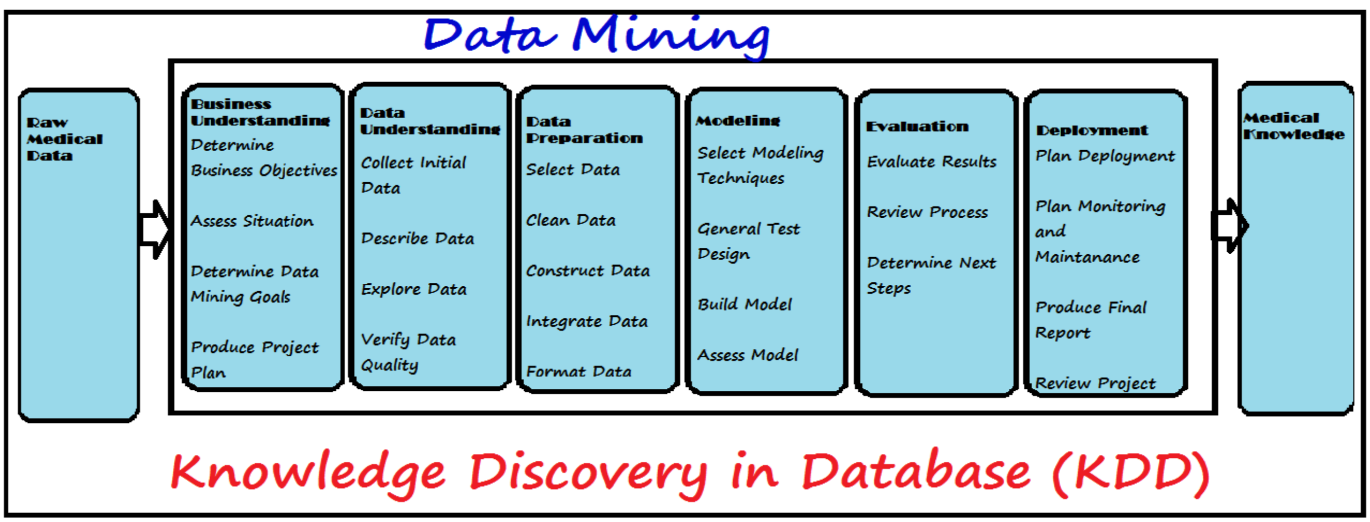


Fig 2.2: Knowledge Discovery

DATA MINING STRATEGIES

The goal of data mining is to learn from data, and there are two broad categories of data mining strategies: supervised and unsupervised learning. Supervised learning methods are deployed (or strategically put into service in an information technology context) when values of variables (inputs) are used to make predictions about another variable (target) with known values.

Unsupervised learning methods can be used in similar situations, but are more frequently deployed on data for which a target with known values does not exist. In supervised methods, the models and attributes are known and are applied to the data to predict and discover information. With unsupervised modeling, the attributes and models of fraud are not known, but the patterns and clusters of data uncovered by data mining can lead to new discoveries (Salim & etal, 2013).

LIMITATIONS OF DATA MINING

Data mining applications can greatly benefit the healthcare industry. However, they are not without limitations.

- Healthcare data mining can be limited by the accessibility of data, because the raw inputs for data mining often exist in different settings and systems, such as administration, clinics, laboratories and more. Hence, the data have to be collected and integrated before data mining can be done.
- While several authors and researchers have suggested that a data warehouse be built before data mining is attempted, that can be a costly and time-consuming.
- Secondly, other data problems may arise whereby this include missing, corrupted, inconsistent or non-standardized data such as pieces of information recorded in different formats in different data sources. In particular, the lack of a standard clinical vocabulary is a serious hindrance to data mining applications.
- Thirdly, there may be ethical, legal and social issues, such as data ownership and privacy issues, related to healthcare data.
- Fourthly, the successful application of data mining requires knowledge of the domain areas as well as in data mining methodology and tools. Without a sufficient knowledge of data mining, the user may not be aware or be able to avoid the pitfalls of data mining (Salim & etal, 2013).

2.5 Methods of demand forecasting in healthcare

According to (Kesten & Armstrong, 2012) Demand forecasters can draw upon many existing forecasting techniques and methods. These methods can be grouped into 17 categories. Twelve rely on judgment, namely unaided judgment, decomposition, expert surveys, 4 structured analogies, game theory, judgmental bootstrapping, intentions and expectations surveys, simulated interaction, conjoint analysis, experimentation, prediction markets, and expert systems. The remaining five methods require quantitative data. They are extrapolation, quantitative analogies, causal models, neural nets, and rule-based forecasting.

While it is beyond the scope of this proposal to present an exhaustive description of forecasting methodologies, the section below highlights a few techniques that require quantitative data and data mining.

Judgmental Forecasts

Judgmental forecasts rely on analysis of such subjective inputs as executive opinions, contracts/insurance/POS company estimates, consumer surveys, mental estimates of the market, intuition, outside (consultant) opinions, and the opinions of managers and staff. A health care manager may use staff to generate a judgmental forecast or several forecasts from which to choose.

Methods requiring quantitative data

Extrapolation

Extrapolation methods require historical data only on the variable to be forecast. They are appropriate when little is known about the factors affecting a variable to be forecast. Statistical extrapolations are cost effective when many forecasts are needed.

Perhaps the most widely used extrapolation method, with the possible exception of using last year's value, is exponential smoothing. Exponential smoothing is sensible in that recent data are weighted more heavily and, as a type of moving average, the procedure smoothes out short-term fluctuations. Exponential smoothing is understandable, inexpensive, and relatively accurate. When extrapolation procedures do not use information about causal factors, uncertainty can be high, especially about the long-term. The proper way to deal with uncertainty is to be conservative. For time series, conservatism requires that estimates of trend be damped toward no change: The greater the uncertainty about the situation, the greater the damping that is needed. Procedures are available to damp the trend and some software packages allow for damping.

One promising extrapolation approach is to decompose time series by causal forces. This is expected to improve accuracy when a time series can be effectively decomposed under two conditions: (1) if domain knowledge can be used to structure the problem so that causal forces differ for two or more component series, and (2) when it is possible to obtain relatively accurate forecasts for each component. For example, to forecast the number of people that will die on the highways each year, forecast the number of passenger miles driven (a series that is expected to grow), and the death rate per million passenger miles (a series expected to decrease), then multiply these forecasts. When tested on five time series that clearly met the conditions, decomposition by causal forces reduced forecast errors by two-thirds. For the four series that partially met the conditions, decomposition by causal forces reduced errors by one-half. Although the gains in accuracy were large, to date there is only the one study on decomposition by causal forces (Armstrong, Collopy and Yokum 2005 in (Kesten & Armstrong, 2012)).

For many years Box-Jenkins was the favored extrapolation procedure among statisticians and it was admired for its rigor. Unfortunately, there are two problems: First, it is difficult for reasonably intelligent human beings to understand. And, second, studies of comparative accuracy found that Box-Jenkins does not improve.

Causal Models

Causal models include models derived using segmentation, regression analysis, and the index method. These methods are useful if knowledge and data are available for variables that might affect the situation of interest. For situations in which large changes are expected, forecasts from causal models are more accurate than forecasts derived from extrapolating the dependent variable. Theory, prior research, and expert domain knowledge provide information about relationships between explanatory variables and the variable to be forecast. Causal models are most useful when;

- strong causal relationships exist,
- the directions of the relationships are known,
- large changes in the causal variables are expected over the forecast horizon, and

- The causal variables can be accurately forecast or controlled, especially with respect to their direction.

Segmentation involves breaking a problem down into independent parts of the same kind, using knowledge and data to make a forecast about each part, and combining the forecasts of the parts. For example, a hardware company could forecast industry sales for each type of product and then add the forecasts.

To forecast using segmentation, identify important causal variables that can be used to define the segments, and their priorities. Determine cut-points for each variable such that the stronger the relationship with the dependent variable, the greater the non-linearity in the relationship, and the more data that are available the more cut-points that should be used. Forecast the population of each segment and the behavior of the population within each segment then combine the population and behavior forecasts for each segment, and sum the segments.

Segmentation has advantages over regression analysis where variables interact, the effects of variables on demand are non-linear, and clear causal priorities exist. Segmentation is especially useful when errors in segment forecasts are likely to be in different directions (Kesten & Armstrong, 2012).

Regression analysis is used to estimate the relationship between a dependent variable and one or more causal variables. Regression is typically used to estimate relationships from historical (non-experimental) data. Regression is likely to be useful in situations in which three or fewer causal variables are important, effect sizes are important, and effect sizes can be estimated from many reliable observations that include data in which the causal variables varied independently of one another (Kesten & Armstrong, 2012).

Important principles for developing regression models are to;

- use prior knowledge and theory, not statistical fit, for selecting variables and for specifying the directions of their effects,

- discard variables if the estimated relationship conflicts with prior evidence on the nature of the relationship
- keep the model simple in terms of the number of equations, number of variables, and the functional form

The *index method* is suitable for situations with little data on the variable to be forecast, where many causal variables are important, and where prior knowledge about the effects of the variables is good. Use prior empirical evidence to identify predictor variables and to assess each variable's directional influence on the outcome. Experimental findings are especially valuable. Better yet, draw on findings from meta-analyses of experimental studies. If prior studies are not available, independent expert judgments can be used to choose the variables and determine the directions of their effects. If prior knowledge on a variable's effect is ambiguous or contradictory, do not include the variable in the model.

The index method is especially useful for selection problems, such as for assessing which geographical location offers the highest demand for a product. The method has been successfully tested for forecasting the outcomes of U.S. presidential elections based on information about candidates' biographies and voters' perceptions of candidates' ability to handle the issues (Kesten & Armstrong, 2012).

Time-Series Approach

A time series is a sequence of evenly spaced observations taken at regular intervals over a period of time (such as daily, hourly, weekly, monthly, or yearly). An example of a time series is the monthly admissions to a multisystem hospital. Forecasts from time-series data assume that future values of the series can be predicted from past values. Analysis of a time series can identify the behavior of the series in terms of trend, seasonality, cycles, irregular variations, or random variations. A trend is a gradual, long-term, upward or downward movement in data. Seasonality refers to short-term, relatively frequent variations generally related to factors such as weather, holidays, and vacations; health care facilities often experience weekly and even daily "seasonal" variations. Cycles are patterns in the data that occur every several years, often in relation to current economic conditions. Such cycles often exhibit wavelike characteristics that mimic the business cycle. Irregular variations are “spikes” in the data caused by chance or unusual circumstances (examples: severe weather, labor strike, use of a new high-technology health service); they do not reflect typical behavior and should be identified and removed from the data whenever possible. Random variations are residual variations that remain after all other behaviors have been accounted for. Graphing the data provides clues to a health care manager for selecting the right forecasting method. Techniques for Averaging Historical data usually contain a certain amount of noise (random variation) that tends to obscure patterns in the data. Randomness arises from a multitude of relatively unimportant factors that cannot possibly be predicted with any certainty.

The optimal situation would be to completely remove randomness from the data and leave only “real” variations (for example, changes in the level of patient demand). Unfortunately, it is usually impossible to distinguish between these two kinds of variations. The best one can hope for is that the small variations are random and the large variations actually mean something. Averaging techniques smooth out some of the fluctuations in a data set; individual highs and lows are “averaged” out. A forecast based on an average shows less variability than the original data set does. The result of using averaging techniques is that minor variations are treated as random variations and essentially “smoothed” out of the data set. Although the larger variations, those deemed likely to reflect “real” changes, are also smoothed, it is to a lesser degree. Three techniques for averaging are described in this section: naive forecasts, moving averages, and exponential smoothing.

Moving Averages (MA). While a naive forecast uses data from the previous period, a moving average forecast uses a number of the most recent actual data values. The moving average forecast is found using the following equation:

EXAMPLE 2.1

$$F_t = \text{MA}_n = \frac{\sum A_i}{n}$$

where

F_t = Forecast for time period t

MA_n = Moving average with n periods

A_i = actual value with age i

i = "age" of the data ($i = 1, 2, 3 \ldots$)

n = number of periods in moving average

An OB/GYN clinic has the following yearly patient visits, and would like to predict the volume of business for the next year for budgeting purposes.

Period (t)	Age	Visits
1	5	15,908
2	4	15,504
3	3	14,272
4	2	13,174
5	1	10,022

Solution: Using formula [2.1], the three-period moving average (MA3) for period 6 is

$$F_6 = \text{MA}_3 = (14{,}272 + 13{,}174 + 10{,}022) \div 3 = 12{,}489.$$

With the available data a health care manager can back forecast earlier periods; this is a useful tool for assessing accuracy of a forecast. Computation of 3-period moving averages for the OB/GYN visits then would look like this:

Period (t)	Age	Visits	Forecast
1	5	15,908	
2	4	15,504	
3	3	14,272	
4	2	13,174	15,228
5	1	10,022	14,317
6			12,489

This technique derives its name from the fact that as each new actual value becomes available, the forecast is updated by adding the newest value and dropping the oldest and then recalculating the average. Thus the forecast "moves" by reflecting only the most recent values. For instance, to calculate the forecasted value of 15,228 for period 4 ($F4$), the visits from periods 1 through 3 were averaged; to calculate $F5$, visits from period 1 were dropped, but visits from period 4 were added to the average. Moving averages based on many data points will be smoother, but less responsive to "real" changes. The decision maker must consider the cost of responding more slowly to changes in the data against the cost of responding to what may be simply random variations (Yasar, 2005).

Neural networks

Neural networks are designed to pick up nonlinear patterns in long time-series. Studies on neural nets have been popular with researchers with more than 7,000 articles identified in an August 2012 Social Science Citation Index (Web of Knowledge) search for the topic of neural networks and forecasting. Early reviews on the accuracy of forecasts from neural nets were not favorable. However, Adya and Collopy (1998) found only eleven studies that met the criteria for a comparative evaluation, and in eight of these, neural net forecasts were more accurate than alternative methods (Kesten & Armstrong, 2012).

Decision trees

Decision trees are powerful and popular tools for classification and prediction. They are so appealing because they are easily understood, as they can be graphically presented as trees as well as in the form of rules (in English or in SQL). Most popular algorithms are CHAID (Chi squared automatic induction), CART and C4.5 (later version of ID3 algorithm). For the purpose of this paper, CHAID is going to be more thoroughly explained. Decision trees are turned upside down and built from the root at the top toward the leaves at the bottom. The nodes of a tree represent questions; an answer to one question determines which question will be asked next. The process starts in the root node, where a record is tested and the result of the test determines lower node where the process will proceed (Mirjana & Dijana, 2008)

CHAID algorithm

CHAID is the oldest algorithm of the mentioned tree, and it was firstly published by Hartigan in 1975. It is derived from the AID (Automatic Interaction Detection) with a purpose to detect statistical relations between variables (which is done by building a decision tree) and it has been used for classification (8). As mentioned above, CHAID is different from other algorithms because it follows prepruning method, that is; it tries to stop the branching before the over fitting (inability to generalize the results based on training set to other data) occurs. Another difference is that CHAID works only with categorical variables so it is necessary to break the ranges in the continued ones, or replace them with classes. The key to CHAID lies in its first two letters that mark the test of significance that is X2 (chi square test). CHAID uses X2 test to make a decision about merging the fields that do not create statistically significant differences in the values of a target field. After that, it splits every group with three or more fields with binary splits, and if some of these splits generate a statistically significant difference in outcomes, CHAID retains that split. Next, X2 test is being used again, and the field that generates the groups that differentiate the most (according to that test) is being chosen as a splitter for that node. The tree is growing until there are no more splits that lead to statistically significant differences in classification. Precise level of significance determines the size of a tree and its value as a classifier (Mirjana & Dijana, 2008).

EVALUATION OF DEMAND FORECASTING TECHNIQUES

The basic purpose of all forecasting techniques is to 'forecast' or measure or assign a value represented by a number to the demand-variable at a future point in time. That is these techniques are supposed to assign the true value of the demand for product, say x, at a future date. This again refers to assigning values to the three components of demand namely the need, the willingness to buy, and the ability to buy i.e. purchasing power (Chendroyaperumal, 2009).

Firstly, do these forecasting techniques do this? All the 'objective' methods treat these three components or variables as one single variable and are trying to assign a single value to represent three different variables! This is where they fundamentally go wrong and this forms the source of weakness or limitations of these techniques. This is purely unrealistic since three properties or variables cannot be represented by a single value. An 'assumption' that a single value represents three properties or variables is as bad as a guess that is untrue and this is totally erroneous. All the subjective methods (consumer surveys, expert opinion surveys) deal with the 'intention' or ‘willingness’ component of the demand whereas the test-market deals with the actual behavior of consumers (i.e. the actual demand) in a test-market and tries to generalize the behavior to similar market conditions (Chendroyaperumal, 2009).

Secondly, will it be possible to measure or forecast these three components of demand scientifically? That is will we be able to arrive at the same value for these variables irrespective of the time, the place, and the person measuring it. The answer to this question raises the nature of the product. All products can be classified into two basic groups namely the 'essentials' and the 'nonessentials'. The essential products are the inputs that are a must for the individual to sustain their functioning or to survive. The non-essential products are those that are not a must for sustaining the functioning or the survival of an individual. The non-essential needs, consisting of quantifiable and no quantifiable, are not finite in number and therefore measuring them accurately is not possible. Any living individual will have the **need** for the 'essentials' with positive willingness to buy them (Chendroyaperumal, 2009).

The **ability to buy** can be measured by the disposable income of an individual in terms of monetary units. There is any number of cases where the individuals have sufficient or huge money, both in the past and in the present, but they do not spend it and hence mere presence of ability-to-buy alone does not cause demand. Therefore using the money income alone as a cause of demand is not realistic. An interesting implication of this definition of demand is that the Western concept of demand, through the absence of the ability-to-buy component of demand, straight away excludes all those human beings without purchasing power from the 'market' and thus ignores unemployment and poverty.

From the above discussion it can be inferred that demand occurs only when all the three components (the need, the willingness, the purchasing power) are present sufficiently and that all these variables are to be 'measured' to forecast demand (Chendroyaperumal, 2009).

2.6 Data mining method of choice in Demand forecasting in healthcare

Neural networks forecasting has proved to be the most accurate than alternative methods and therefore is our method of choice. The schematic diagram of a Neural Network is show in the figure below.

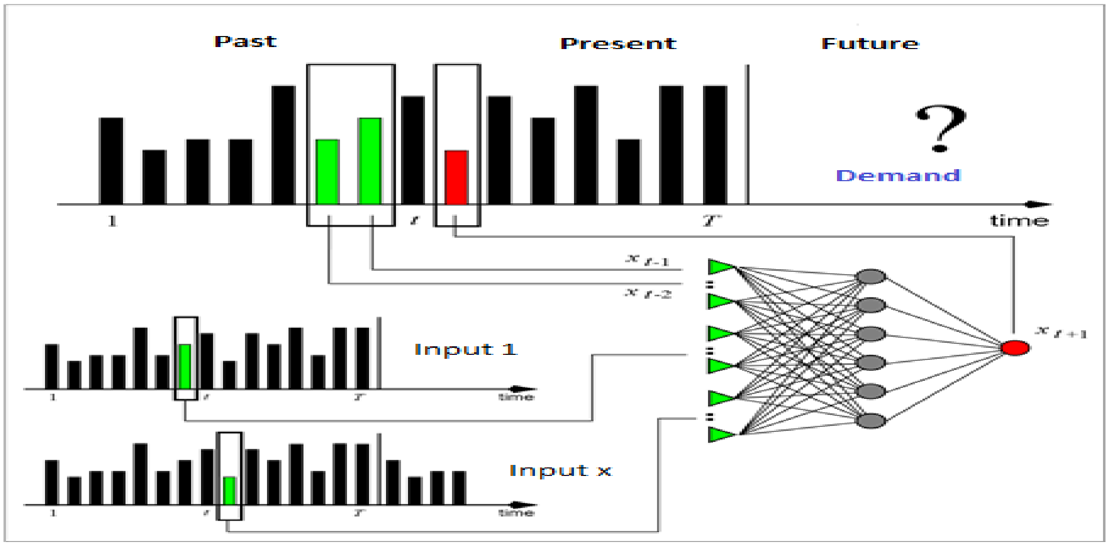


Figure 2.3: Data Mining using Neural Networks

The neuron accepts several inputs and gives out one output. There are hidden layers that the inputs pass through to generate output.

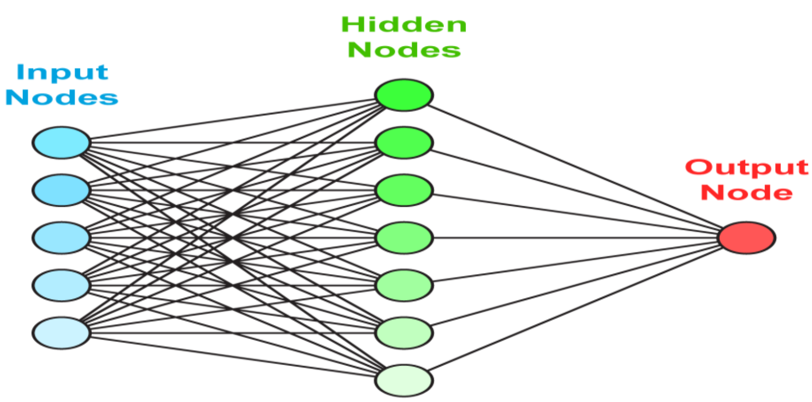


Figure 2.4: Hidden layers of a Neural Networks

An Artificial Neural Network (ANN) is a system based on the operation of biological neural networks. A neural network can perform tasks that a linear program cannot; when an element of the neural network fails, it can continue without any problem by their parallel nature; a neural network learns and does not need to be reprogrammed. The input/output training data are fundamental in neural network technology, because they convey the necessary information to "discover" the optimal operating point. The key factors that make ANN highly suitable for demand forecasting include: it does not require any pre-assumed functional relationship between the factors of demand like Income, Population and Literacy levels (Moturi & Kioko, 2013).

CHAPTER 3: METHODOLOGY

This section presents a detailed description of the selected research design. It describes in detail what is to be done and how it will be done.

3.1 Research Site

Nairobi city is the commercial capital of Kenya and is the most populous city in East Africa and is currently the 12th largest city in Africa. According to the 2009 Census of Kenya, in the administrative area of Nairobi city, 3,138,295 inhabitants lived within 696 km2 (269 sq. mi). The city is politically administered by the county government and is sub-divided into constituencies. There are two main classes of health facilities at this city. They are called public and private health facilities. The former covers health centers and hospitals owned by the Ministry of Health while the latter covers privately owned health care institutions.

A number of factors were considered in selecting this county as the study area; These includes accessibility to health demand data and the types of health services that are available at the healthcare facilities. In addition, all the planning issues that are dealt with in this county are relevant to the remaining hospitals in the country.

3.2 Research Design

Data mining is a creative process which requires a number of different skills, knowledge and tools. It needs a standard approach which will help translate business problems into data mining tasks, suggest appropriate data transformations and data mining techniques, and provide means for evaluating the effectiveness of the results and documenting the experience (Rudiger & Jochen, 2000).

The CRISP-DM (CRoss Industry Standard Process for Data Mining) defined a process model which provides a framework for carrying out data mining projects which is independent of both the industry sector and the technology used. The CRISP-DM process model aims to make large data mining projects, less costly, more reliable, more repeatable, more manageable, and faster. This is the methodology of choice in this project (Rudiger & Jochen, 2000).

The Generic CRISP-DM Reference Model

The CRISP-DM reference model for data mining provides an overview of the life cycle of a data mining project. It contains the phases of a project, their respective tasks, and their outputs. The life cycle of a data mining project is broken down in six phases which are shown in figure below. The arrows indicate only the most important and frequent dependencies between phases, but in a particular project, it depends on the outcome of each phase which phase, or which particular task of a phase, has to be performed next.

The outer circle in figure symbolizes the cyclic nature of data mining itself. Data mining is not finished once a solution is deployed. The lessons learned during the process and from the deployed solution can trigger new, often more focused business questions. Subsequent data mining processes will benefit from the experiences of previous ones (Rudiger & Jochen, 2000).

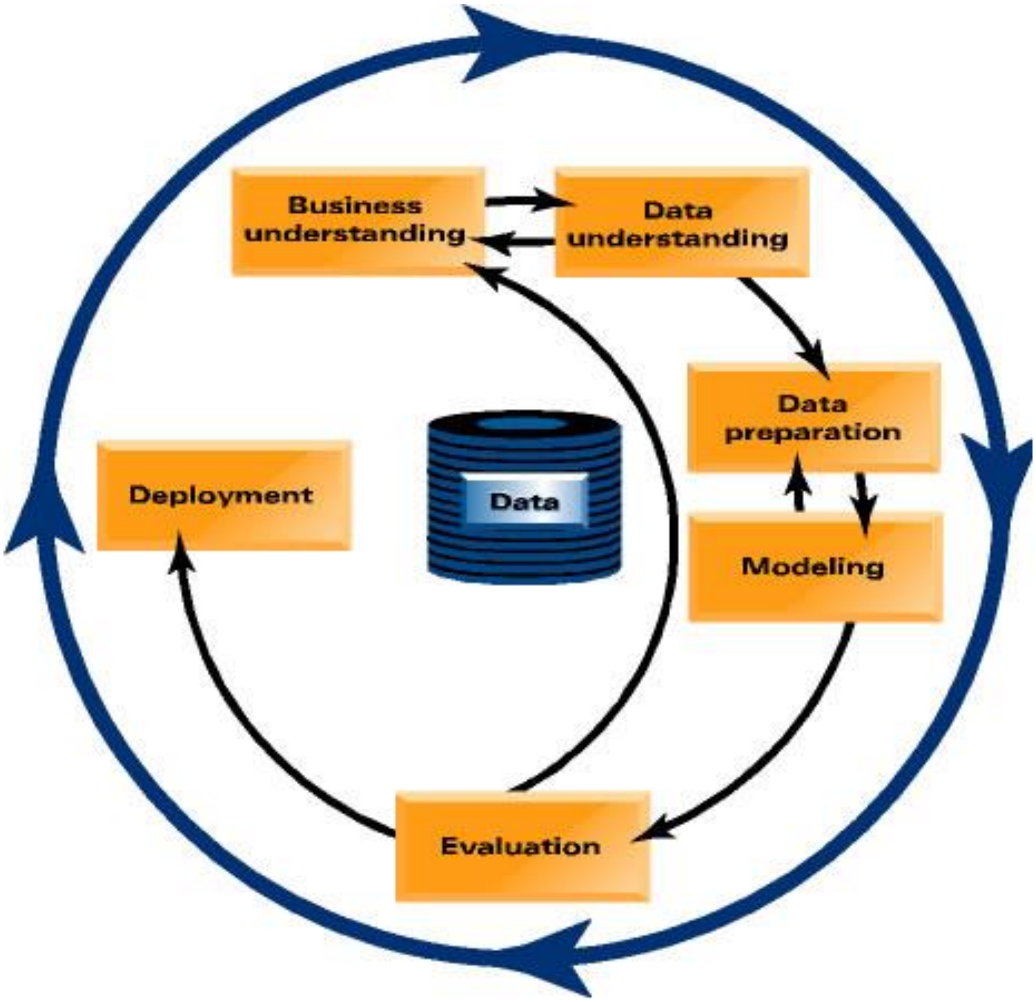


Figure 3.1: The life cycle of a data mining project

In the following, I outline each phase briefly:

Business Understanding

This initial phase focuses on understanding the project objectives and requirements from a business perspective, and then converting this knowledge into a data mining problem definition, and a preliminary project plan designed to achieve the objectives (Rudiger & Jochen, 2000).

- Understand the business objectives
- Current Systems Assessment
- Task Decomposition

- Identify Constraints
- Build a project plan

Data Understanding

The data understanding phase starts with an initial data collection and proceeds with activities in order to get familiar with the data, to identify data quality problems, to discover first insights into the data, or to detect interesting subsets to form hypotheses for hidden information. There is a close link between Business Understanding and Data Understanding. The formulation of the data mining problem and the project plan require at least some understanding of the available data (Rudiger & Jochen, 2000).

- Collect Data
- Data Description
- Data Exploration

Data Preparation

The data preparation phase covers all activities to construct the final dataset (data that will be fed into the modeling tool(s)) from the initial raw data. Data preparation tasks are likely to be performed multiple times, and not in any prescribed order. Tasks include table, record, and attribute selection, data cleaning, construction of new attributes, and transformation of data for modeling tools.

- Integrate Data
- Select Data
- Data Transformation
- Clean Data
- Data Construction

Modeling

In this phase, various modeling techniques are selected and applied, and their parameters are calibrated to optimal values. Typically, there are several techniques for the same data mining problem type. Some techniques require specific data formats. There is a close link between Data Preparation and Modeling. Often, one realizes data problems while modeling or one gets ideas for constructing new data.

- Select of the appropriate modelling technique

- Develop a testing regime
- Build Model
- Assess the model

Evaluation

At this stage in the project you have built one or more models that appear to have high quality, from a data analysis perspective. Before proceeding to final deployment of the model, it is important to more thoroughly evaluate the model, and review the steps executed to construct the model, to be certain it properly achieves the business objectives. A key objective is to determine if there is some important business issue that has not been sufficiently considered. At the end of this phase, a decision on the use of the data mining results should be reached.

- Validate Model
- Review Process
- Determine next steps

Deployment

Creation of the model is generally not the end of the project. Usually, the knowledge gained will need to be organized and presented in a way that the customer can use it. Depending on the requirements, the deployment phase can be as simple as generating a report or as complex as implementing a repeatable data mining process. In many cases it will be the user, not the data analyst, who will carry out the deployment steps. In any case, it is important to understand up front what actions will need to be carried out in order to actually make use of the created models.

- Knowledge Deployment is specific to objectives
- Process deployment/production
- Produce final project report

Overview of WEKA

This is an acronym for Waikato Environment for Knowledge Analysis which is a free and open source software used for machine learning algorithms and data mining tasks. WEKA implements different algorithms which include data pre-processing, classification, Decision Trees, Artificial Neural Networks, data analysis, predictive modeling and logic regression as well as contains a collection of visualization tools. This will be the software tool of choice in this project

How WEKA Works

WEKA allows the GUI user to select the four different ways to work with. These four ways are ; Explorer, Experimenter, KnowledgeFlow or a simple CLI. In this project, Explorer will be the method of choise to work with.

WEKA accepts data in ARFF (Attribute-Relation File Format) formats which is an ASCII text file that describes a list of instances sharing a set of attributes. These files have two sections: Header information sector and Data information sector. The header contains the name of the relation, a list of attributes and their data types. However using JDBC, users can access data from Windows Databases like MS Access or other Databases in general (MySQL, PostgreSQL, MS SQL Server, Oracle, etc.) from WEKA. WEKA can be used from any three stated below interfaces; the command line, Graphical User Interface, from java programs

3.3 Study Population and sources of Data

This study will focus on forecasting of Health Services demand in Nairobi County. The study will rely on secondary data collected by various players in the industry for their specific use.

Data requirements will include;

a. Nairobi county distribution of Level 4-6 hospitals
b. Nairobi county current population in constituencies from Kenya National bureau of statistics
c. Nairobi county hospital visits from e-Health
d. Nairobi county diseases diagnosed and surgeries done from the ministry of Health
e. Average Income per person per Constituency in the county
f. Population growth rate in the country/county
g. Nairobi County shape files for GIS mapping

3.4 Data Collection Methods

Data collection methods employed will depend on;

- Credibility of findings
- Staff skills
- Costs
- Time constraints

The various methods to be used includes;

1. Interviews

   This approach will be applied to gather in-depth attitudes, beliefs, and anecdotal data from individual players in the industry. It will require researcher time and quiet area to conduct interviews and equipment to record and transcribe interviews.

   In-depth interviews in which the interviewer does not follow a rigid form of questions will be used. The interviewer will seek to encourage free and open responses, and there may be a tradeoff between comprehensive coverage of topics and in-depth exploration of a more limited set of questions. It interviews will also encourage capturing respondents' perceptions in their own words, a very desirable strategy in qualitative data collection. This will allow the evaluator to present the meaningfulness of the experience from the respondent's perspective. This will be applied to medical personnel in the industry

2. Observations

   Observational techniques are methods by which an individual or individuals gather firsthand data on programs, processes, or behaviors being studied. They provide evaluators with an opportunity to collect data on a wide range of behaviors, to capture a great variety of interactions, and to openly explore the evaluation topic.

3. Document Studies

Existing records will provide insights into a setting and/or group of people that cannot be observed or noted in another way. This information will be found in document form. The usefulness of existing sources will vary depending on whether they are accessible and accurate.

3.5 Data Processing and Analysis

This will start in the field, with checking for completeness of the data and performing quality control checks, while sorting the data by group of informants. The plan will involve;

- Sorting data
- Performing quality-control checks
- Data processing
- Data analysis

3.6 Ethical Considerations

Privacy and confidentiality

The data involved in the study is mainly private medical information. This will need to be handled with ultimate confidentiality for the purpose of this study only.

3.7 Expected Deliverables

Table 3.1: Objectives and Deliverables

Objectives	Deliverables
i. To explore the various forecasting methods in healthcare services demand	In the Literature review, a number of demand forecasting methods have been highlighted and how they work.
ii. To identify the most suitable healthcare services demand forecasting methods for Nairobi County.	The most accurate one as viewed by different authors is use of Artificial Neural Networks compared to Linear Regression and Decision trees.
iii. To develop and validate a data mining model for forecast of private healthcare services demand for Nairobi County.	By applying CRISP-DM methodology and WEKA and the software tool, I am to build a model and validate it. I shall implement

	the best results from the different algorithms.

3.8 Designing the Proposed Model

3.8.1 Requirement Analysis

Business Understanding

To develop and validate a data mining model for forecast of private healthcare services demand for Nairobi County was the key objective of this research. From the current assessment, there does not exists a system that uses the data mining algorithms to forecast health services demand in Nairobi County. Forecasts would enable the managers to anticipate the future demand and plan accordingly.

This project heavily employed use of data mining approach to build a healthcare services demand forecasting model in Nairobi County as a solution to this problem. This demand forecasting model will give the health service providers, managers and planners an optimal decision they can make today about the private healthcare business tomorrow. The main constrains in the project was getting clean data to be able to build the model. However, the data that was finally collected is authentic and reliable and so is the model results.

Factors of Forecast

Demand Forecast for health services is a complete analysis of inpatient and emergency department outpatient volume growth projections in Nairobi County over the next ten years. The analysis addressed key questions facing today's hospital leaders and health planners.

- What kind of services will be demanded by the residents, and at what volume?
- Where are the pockets of growth in the market?

Why forecast?

1. Capacity planning
    - Number and type of beds
    - Surgical suites

- Equipment

2. Analyze opportunities – market share by various players
3. Revenue projections by the hospitals
4. Assessment of medical staff
    - Physician succession and recruitment
    - Nursing staff
    - Research
5. Service area demographics/population shifts
6. Utilization trends
7. Medical technologies/products/pharmaceutical pipeline

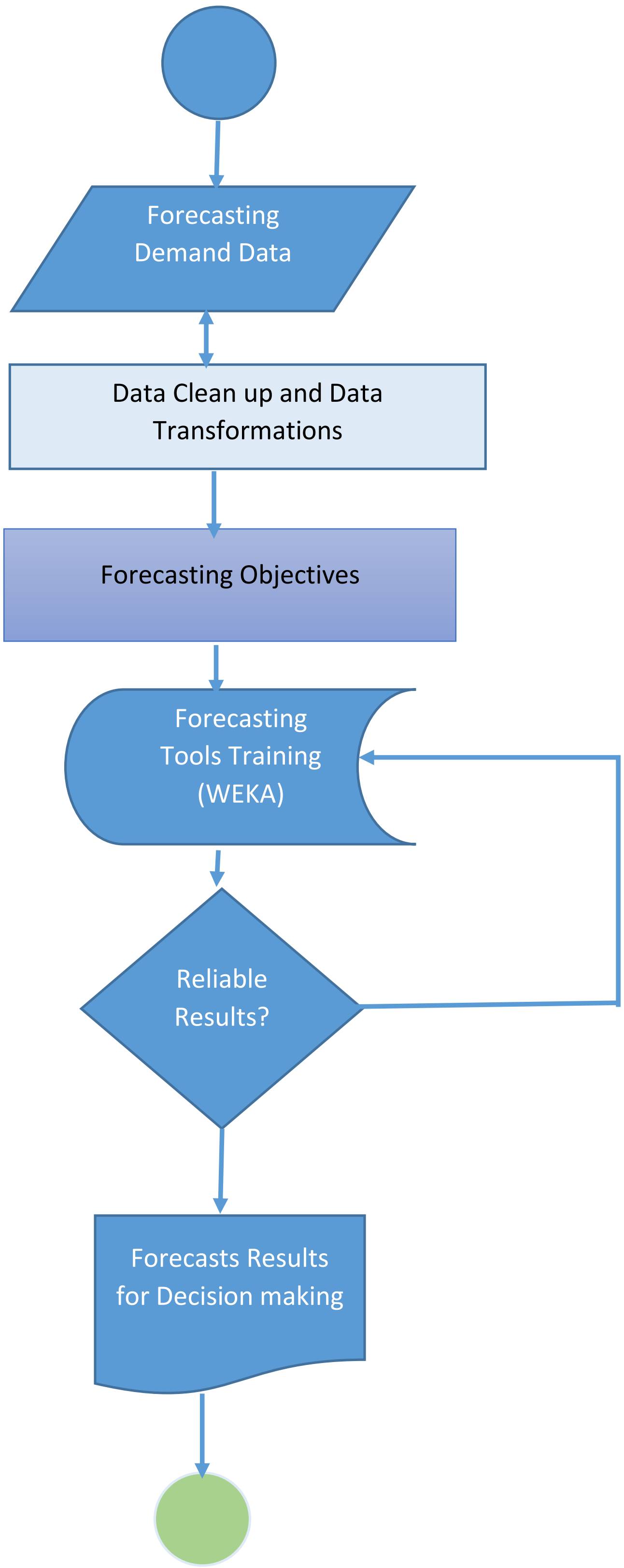


Figure 3.2: Forecasting Process Flow diagram

3.8.2 Architectural Design

This section presents the overall architectural design of the Demand Forecasting system basing it on the various subsystems that will be interconnected by means of control flows and data.

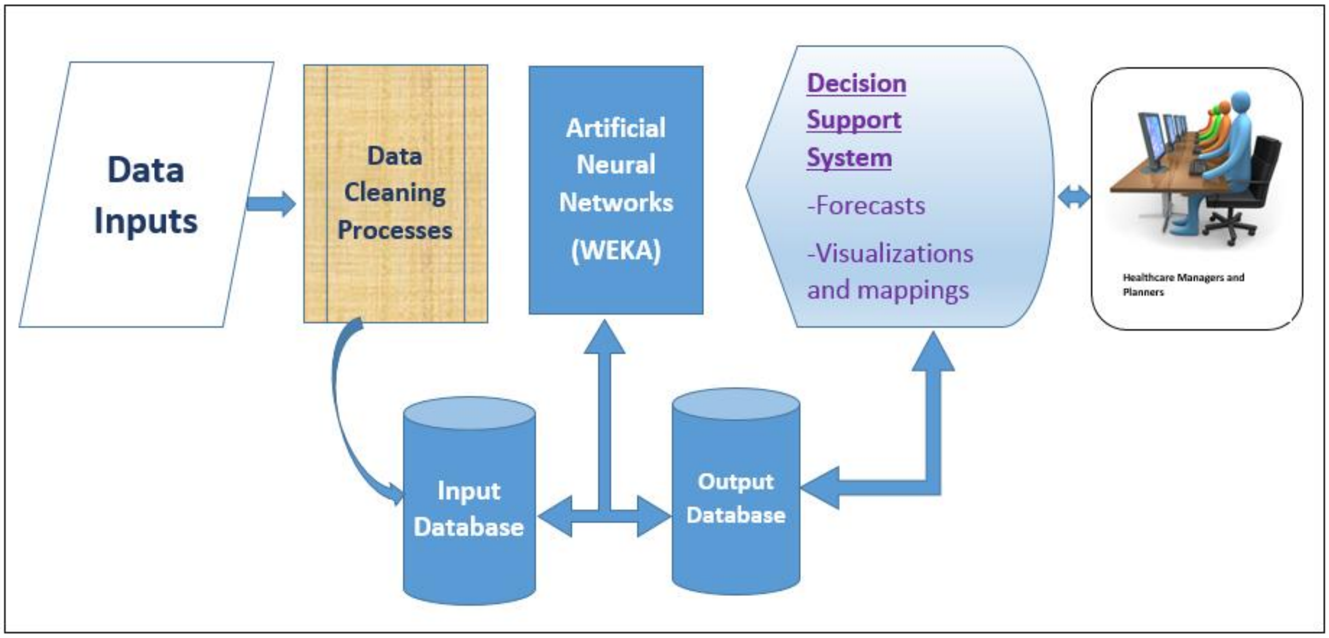


Figure 3.3: Design architecture

3.8.3 Data Understanding

The data set used for the building of the model was collected through the extraction, formatting and organization of the health services data in Nairobi County and its sub counties from Kenya Health Information System. The data collected included all the inpatients, outpatients and surgeries one in all health facilities in the county totaling to more than 1000 facilities.

1	Total AL R	Total Entr	Total Injury eye	Total Retinoblastoma	Total Purulent Conjunctivitis	Total deliveries	Turnover per Bed	Total Outpatient attendance	II
38	15407	0	0	0	0	12255	38.1	328389	
39	6427	0	28	0	35	14592	42.2	801659	
40	8670	0	0	0	0	5969	23.9	790787	
41									
42	2814	0	0	0	0	1899	20.6	599420	
43									
44	540	0	0	0	0	1505	11.8	397666	
45	1994	0	0	0	0	9347	95.3	537372	
46	31206	5	20	2	83	4712	88.2	365029	
47	840	0	154	0	354	8234	63.2	746292	
48	99	0	0	0	0	13689	17.1	274055	
49	30302	0	0	0	0	7439	50.3	821594	
50	103	0	0	0	0	7297	27.2	485182	
51	435	0	0	0	0	3635	28.7	521115	
52									
53	401	0	0	0	0	910	46	305919	
54									
55	1306	0	0	0	0	1029	12.9	301448	
56	1440	0	0	0	0	6257	86.8	379440	
57									
58									
59									

data-8 | Data Elements Summary | GDP | OP | Error Magins | OP Specific Cases | Surgeries

Figure 3.4: Outpatients raw data

More data was also collected for the health facilities in the county with their levels or classifications, ownership, bed capacity and location, this from eHealth Kenya. This included more than 9000 facilities in Kenya and was then filtered to work with the facilities in Nairobi County.

More data was also collected for the health facilities in the county with their levels or classifications, ownership, bed capacity and location, this from eHealth Kenya. This included more than 9000 facilities in Kenya and was then filtered to work with the facilities in Nairobi County.

1	Facility Code	Facility Name	Type	Owner	Province	Constituency	KEPH Level	Beds
281	13213	St Lukes (Kona) Health Centre	Health Centre	Private Enterprise (Institution)	Nairobi	DAGORETI	Level 4	7
282	17411	Mama Lucy Kibaki Hospital - Emb	District Hospital	Ministry of Health	Nairobi	EMBAKASI	Level 4	112
283	13257	Wentworth Hospital	Other Hospital	Private Enterprise (Institution)	Nairobi	EMBAKASI	Level 4	12
284	13095	Moi Air Base Hospital	Other Hospital	Ministry of Health	Nairobi	KAMUKUNJI	Level 4	
285	12865	Afwan Medical Center	Nursing Home	Private Enterprise (Institution)	Nairobi	KAMUKUNJI	Level 4	40
286	13034	Kilimanjaro Nursing Home	Nursing Home	Private Enterprise (Institution)	Nairobi	KAMUKUNJI	Level 4	11
287	13157	Pumwani Maternity VCT Centre	VCT Centre (Stand-Alone	Local Authority	Nairobi	KAMUKUNJI	Level 4	350
288	13182	Shaam Nursing Home	Nursing Home	Private Enterprise (Institution)	Nairobi	KAMUKUNJI	Level 4	10
289	13000	Kamiti Prison Hospital	Other Hospital	Ministry of Health	Nairobi	KASARANI	Level 4	
290	17936	ST JOHN HOSPITAL LIMITED	Other Hospital	Private Practice - Unspecified	Nairobi	KASARANI	Level 4	42
291	13042	Langata Hospital	Other Hospital	Private Enterprise (Institution)	Nairobi	LANG'ATA	Level 4	15
292	13109	Meridian Equator Hospital	Other Hospital	Private Enterprise (Institution)	Nairobi	LANG'ATA	Level 4	54
293	13115	Nairobi West Hospital	Other Hospital	Private Enterprise (Institution)	Nairobi	LANG'ATA	Level 4	110
294	13218	St Mary's Mission Hospital	Other Hospital	Kenya Episcopal Conference-Catholic S	Nairobi	LANG'ATA	Level 4	280
295	17391	St.Mac's Hospital	Other Hospital	Community	Nairobi	LANG'ATA	Level 4	4
296	12984	Jamaa Mission Hospital	Other Hospital	Kenya Episcopal Conference-Catholic S	Nairobi	MAKADARA	Level 4	65
297	13074	The Mater Hospital Mukuru	Other Hospital	Kenya Episcopal Conference-Catholic S	Nairobi	MAKADARA	Level 4	120
298	13076	Mathari Hospital	District Hospital	Ministry of Health	Nairobi	STAREHE	Level 4	700
299	12965	Guru Nanak Hospital	Other Hospital	Private Enterprise (Institution)	Nairobi	STAREHE	Level 4	38
300	12905	Coptic Hospital(Ngong Road)	Other Hospital	Christian Health Association of Kenya	Nairobi	WESTLANDS	Level 4	60
301	13001	Kangemi Health Centre	Health Centre	Local Authority	Nairobi	WESTLANDS	Level 4	20
302	13062	Maria Immaculate Health Centre	Health Centre	Kenya Episcopal Conference-Catholic S	Nairobi	WESTLANDS	Level 4	27
303	13201	St Florence Medical Care Health C	Health Centre	Private Enterprise (Institution)	Nairobi	WESTLANDS	Level 4	12
304	13087	Memorial Hospital	Other Hospital	Armed Forces	Nairobi	DAGORETI	Level 5	
305	13156	Pumwani Maternity Hospital	Other Hospital	Local Authority	Nairobi	KAMUKUNJI	Level 5	350
306	13004	The Karen Hospital	Other Hospital	Private Enterprise (Institution)	Nairobi	LANG'ATA	Level 5	102
307	12867	Aga Khan Hospital	Other Hospital	Private Enterprise (Institution)	Nairobi	WESTLANDS	Level 5	600
308	12874	Avenue Nursing Hospital	Other Hospital	Private Enterprise (Institution)	Nairobi	WESTLANDS	Level 5	160
309	12950	Gertrudes Hospital	Other Hospital	Private Enterprise (Institution)	Nairobi	WESTLANDS	Level 5	85
310	13098	MP Shah Hospital	Other Hospital	Private Enterprise (Institution)	Nairobi	WESTLANDS	Level 5	
311	13110	Nairobi Hospital	Other Hospital	Private Enterprise (Institution)	Nairobi	WESTLANDS	Level 5	
312	13117	Nairobi Womens Hospital	Other Hospital	Private Enterprise (Institution)	Nairobi	WESTLANDS	Level 5	8
313	13023	Kenyatta National Hospital	National Referral Hospita	State Coorporation	Nairobi	DAGORETI	Level 6	1455
314	13194	National Spinal Injury Hospital	National Referral Hospita	Ministry of Health	Nairobi	WESTLANDS	Level 6	30
315	12866	Afya Medical Health Centre	Health Centre	Private Enterprise (Institution)	Nairobi	DAGORETI	Not Classified	0
316	12938	Family Health Medical Dispensar	Dispensary	Non-Governmental Organizations	Nairobi	DAGORETI	Not Classified	0

Facilities | Nairobi Health Facilities

Figure 3.5: Medical Facilities Data

3.8.4 Data Preparation

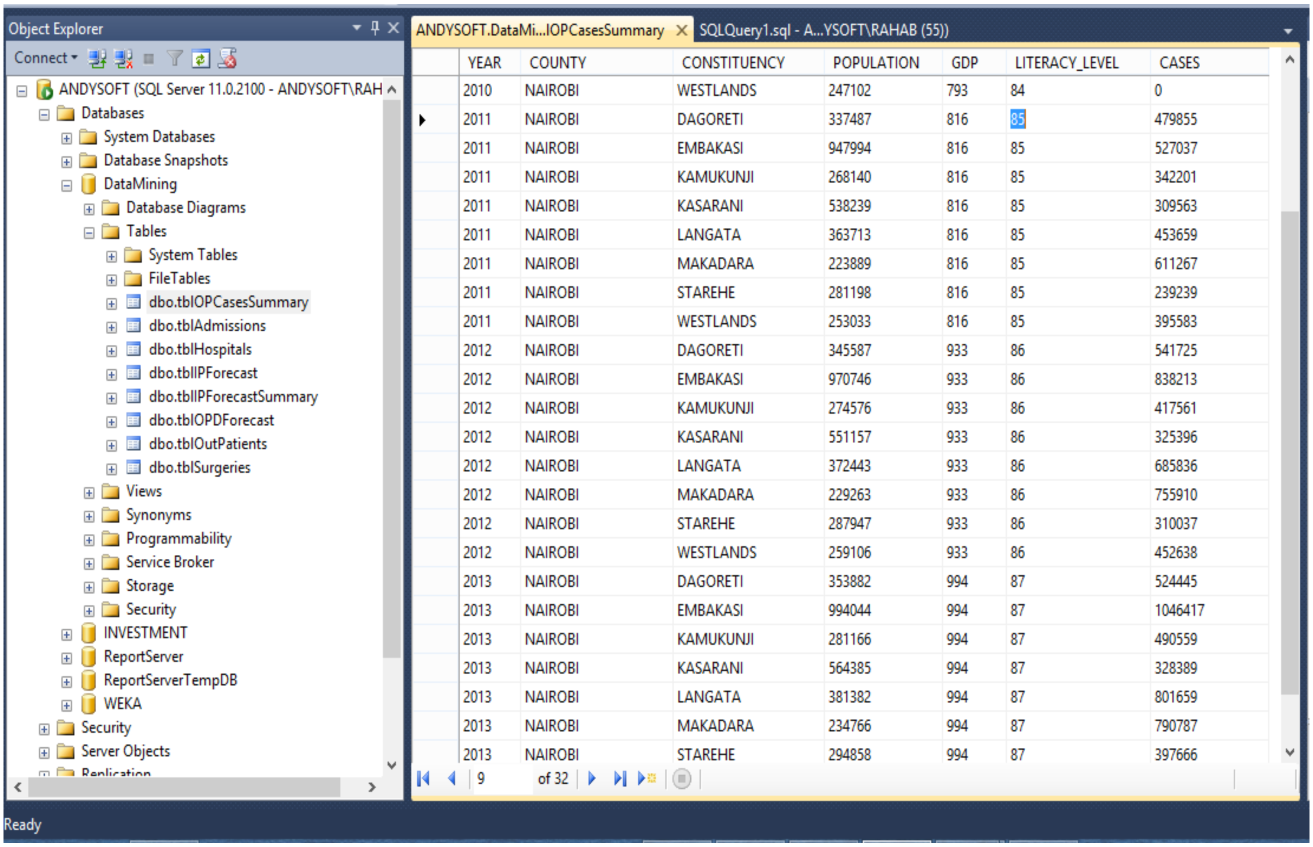
The data collected was very comprehensive and detailed containing a lot of information that was not going to be used to build the model. It included records for and from all the health facilities in the country. This data therefore needed to be filtered and required attributes selected for the records that were in the context of this research project. Missing attributes were reconstructed using lookup mechanisms from the files that had the data or filling using default values. The columns were also labeled appropriately in readiness for uploads in the database.

1	YEAR	COUNTY	CONSTITUENCY	POPULATION	GDP	LITERACY_LEVEL	CASES
2	2010	NAIROBI	DAGORETI	329577	793	84	0
3	2010	NAIROBI	EMBAKASI	925775	793	84	0
4	2010	NAIROBI	KAMUKUNJI	261855	793	84	0
5	2010	NAIROBI	KASARANI	525624	793	84	0
6	2010	NAIROBI	LANGATA	355188	793	84	0
7	2010	NAIROBI	MAKADARA	218641	793	84	0
8	2010	NAIROBI	STAREHE	274607	793	84	0
9	2010	NAIROBI	WESTLANDS	247102	793	84	0
10	2011	NAIROBI	DAGORETI	337487	816	85	479855
11	2011	NAIROBI	EMBAKASI	947994	816	85	527037
12	2011	NAIROBI	KAMUKUNJI	268140	816	85	342201
13	2011	NAIROBI	KASARANI	538239	816	85	309563
14	2011	NAIROBI	LANGATA	363713	816	85	453659
15	2011	NAIROBI	MAKADARA	223889	816	85	611267
16	2011	NAIROBI	STAREHE	281198	816	85	239239
17	2011	NAIROBI	WESTLANDS	253033	816	85	395583
18	2012	NAIROBI	DAGORETI	345587	933	86	541725
19	2012	NAIROBI	EMBAKASI	970746	933	86	838213
20	2012	NAIROBI	KAMUKUNJI	274576	933	86	417561
21	2012	NAIROBI	KASARANI	551157	933	86	325396
22	2012	NAIROBI	LANGATA	372443	933	86	685836
23	2012	NAIROBI	MAKADARA	229263	933	86	755910

data-8 | Data Elements Summary | OP | GDP

Figure 3.6: Formatted Outpatients data

After the transformation process, the data collected was upload in the model Database developed in Microsoft SQL Server database management system.



YEAR	COUNTY	CONSTITUENCY	POPULATION	GDP	LITERACY_LEVEL	CASES
2010	NAIROBI	WESTLANDS	247102	793	84	0
2011	NAIROBI	DAGORETI	337487	816	85	479855
2011	NAIROBI	EMBAKASI	947994	816	85	527037
2011	NAIROBI	KAMUKUNJI	268140	816	85	342201
2011	NAIROBI	KASARANI	538239	816	85	309563
2011	NAIROBI	LANGATA	363713	816	85	453659
2011	NAIROBI	MAKADARA	223889	816	85	611267
2011	NAIROBI	STAREHE	281198	816	85	239239
2011	NAIROBI	WESTLANDS	253033	816	85	395583
2012	NAIROBI	DAGORETI	345587	933	86	541725
2012	NAIROBI	EMBAKASI	970746	933	86	838213
2012	NAIROBI	KAMUKUNJI	274576	933	86	417561
2012	NAIROBI	KASARANI	551157	933	86	325396
2012	NAIROBI	LANGATA	372443	933	86	685836
2012	NAIROBI	MAKADARA	229263	933	86	755910
2012	NAIROBI	STAREHE	287947	933	86	310037
2012	NAIROBI	WESTLANDS	259106	933	86	452638
2013	NAIROBI	DAGORETI	353882	994	87	524445
2013	NAIROBI	EMBAKASI	994044	994	87	1046417
2013	NAIROBI	KAMUKUNJI	281166	994	87	490559
2013	NAIROBI	KASARANI	564385	994	87	328389
2013	NAIROBI	LANGATA	381382	994	87	801659
2013	NAIROBI	MAKADARA	234766	994	87	790787
2013	NAIROBI	STAREHE	294858	994	87	397666

Figure 3.7: Loaded data in SQL Server Database

3.8.5 Designing the Baseline Forecasting Model

Artificial Neural Networks

An Artificial Neural Network (ANN) is a system based on the operation of biological neural networks. The model is founded upon the functionality of a biological neuron. Software implementation of a neural network can be made with their advantages and disadvantages. The advantages include: A neural network can perform tasks that a linear program cannot; when an element of the neural network fails, it can continue without any problem by their parallel nature; a neural network learns and does not need to be reprogrammed; can be implemented in any application; and can be implemented without any problem. The disadvantages include: neural network needs training to operate; architecture of a neural network is different from the architecture of microprocessors therefore needs to be emulated; requires high processing time for large neural networks. ANN is an adaptive, most often nonlinear system that learns to perform a function from data. ANN is built with a systematic step-by-step procedure to optimize a performance criterion or to follow some implicit internal constraint, which is commonly referred to as the learning rule. The input/output training data are fundamental in neural network technology, because they convey the necessary information to "discover" the optimal operating point. The nonlinear nature of the neural network processing elements provides the system with lots of flexibility to achieve practically any desired input/output map.

The key factors that make ANN highly suitable for demand forecasting include: it does not require any pre-assumed functional relationship between the factors of demand like Income, Population and Literacy levels (Moturi & Kioko, 2013).

Training of Artificial Neural Networks

ANN has to be configured such that the application of a set of inputs produces the desired set of outputs. Various methods to set the strengths of the connections exist. One way is to set the weights explicitly, using a priori knowledge. Another way is to 'train' the neural network by feeding it teaching patterns and letting it change its weights according to some learning rule. The learning situations can be categorized in two distinct sorts: *Supervised Learning* or Associative Learning in which the network is trained by providing it with input and matching output patterns ((Moturi & Kioko, 2013). In this research and model building, we used a supervised ANN model using the actual demand data of the Kenya DHIS2 system.

Model Training

The model was trained by applying all the collected data inputs to the network before the weights and biases are updated (batch training mode). The model was then trained using the incremental training mode (i.e. the weights and biases of the network are iteratively adjusted every time an input is applied to the network). The demand forecasted by the WEKA was compared to the actual demand data and the error was calculated using the statistical indices: Mean Percentage Error – to measure how close the forecasts were to the actual load.

Mean Percentage Error

The mean percentage error (MPE) is the computed average of percentage errors by which forecasts of a model differ from actual values of the quantity being forecast. The formula for the mean percentage error is

$$\mathrm{MPE} = \frac{100\%}{n} \sum_{t=1}^{n} \frac{f_t - a_t}{a_t}$$

(Source: Wikipedia)

where $\boldsymbol{a_t}$ is the actual value of the quantity being forecast, $\boldsymbol{f_t}$ is the forecast, and $\boldsymbol{n}$ is the number of different times for which the variable is forecast.

Using the data collected from the sources between 2011 and 2013 and all the values of the factors of demand, the baseline trained model was used to forecast already know data for the year 2013. The results were as shown below.

Table 3.2: WEKA OP 2013 forecasts against the Actual

YEAR	CONSTITUENCY	OP CASES ACTUAL	WEKA OP FORECAST	WEKA % ERROR
2013	DAGORETTI	524445	563440	7.435479412
2013	EMBAKASI	1046417	1078609	3.07640262
2013	KAMUKUNJI	490559	502732	2.48145483
2013	KASARANI	328389	341189	3.897816309
2013	LANGATA	801659	891215	11.17133345
2013	MAKADARA	790787	830190	4.982757683
2013	STAREHE	397666	416699	4.786177345
2013	WESTLANDS	537372	509550	-5.177418995
	MEAN PERCENTAGE ERROR (MPE)			**4.081750331**

The MPE was at 4.08%

3.9 Implementation of the Model

3.9.1 Database implementation

To perform the demand forecasts, several database tables were created in using SQL Server 2012 Database Management System. This database has several tables including the following:

1) tblAdmission – Has raw data on all hospital admissions and inpatients
2) tblHospitals – has information on all the private hospital on whose data forecasting is done.
3) tblIPForecast – Used for forecasting summaries
4) tblIPForecastSummary – used for storing forecast results from WEKA
5) tblOutpatients - used to store raw data on all the outpatient cases
6) tblSurgeries - Stores data on all surgeries.

3.9.2 Development

The Health Demand forecast was done using data mining and forecasting based on WEKA libraries. The model front end was implemented using Visual Basic.NET 2012 and .NET class libraries. Database connectivity was implemented through ODBC. This was the same connection used to read data from the database into WEKA for forecasting. The Reporting in the front end Decision Support system was implemented using SAP Crystal Reports for Visual Studio 2012.

3.9.3 The Prototype's Graphical User Interface (GUI)

For powerful implementation of the Decision Support front end system, the GUI was implemented using Windows forms in Visual Studio 2012. The interface is used for user interactions and has the following functionalities;

1. Inputting, viewing and querying of master static data
    a. Hospitals
    b. Constituencies
    c. Decision Support parameters
2. Loading and viewing raw data used for forecasting
    a. Inpatients data
    b. Outpatients ailments data
    c. Surgeries
3. Inputting of Forecasting results from WEKA
    a. Forecasts on Inpatients
    b. Forecasts on outpatients ailments
    c. Forecasts on Surgeries

4. Decision Support Function
    a. Visualization of Forecasts in Graphs to the user
    b. System suggestions on what decision to make given the forecasts and decision support parameters.

Below are screenshot that shows some of the Windows that implement the above functionalities;

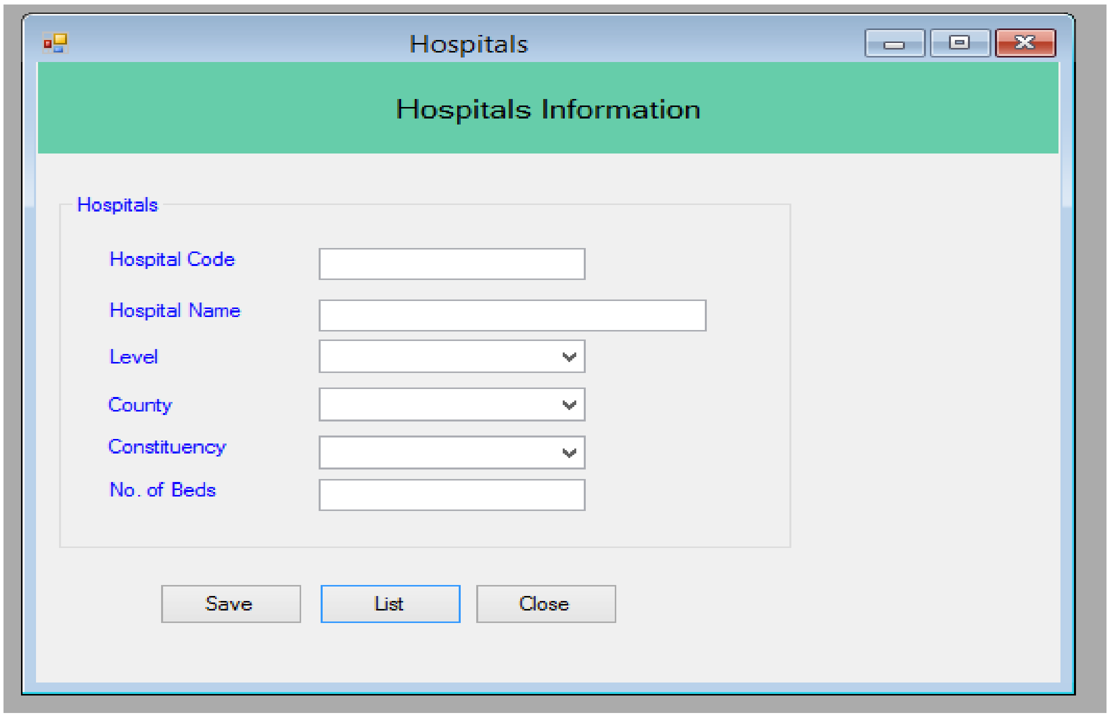


Figure 3.8: Hospitals list screen shot

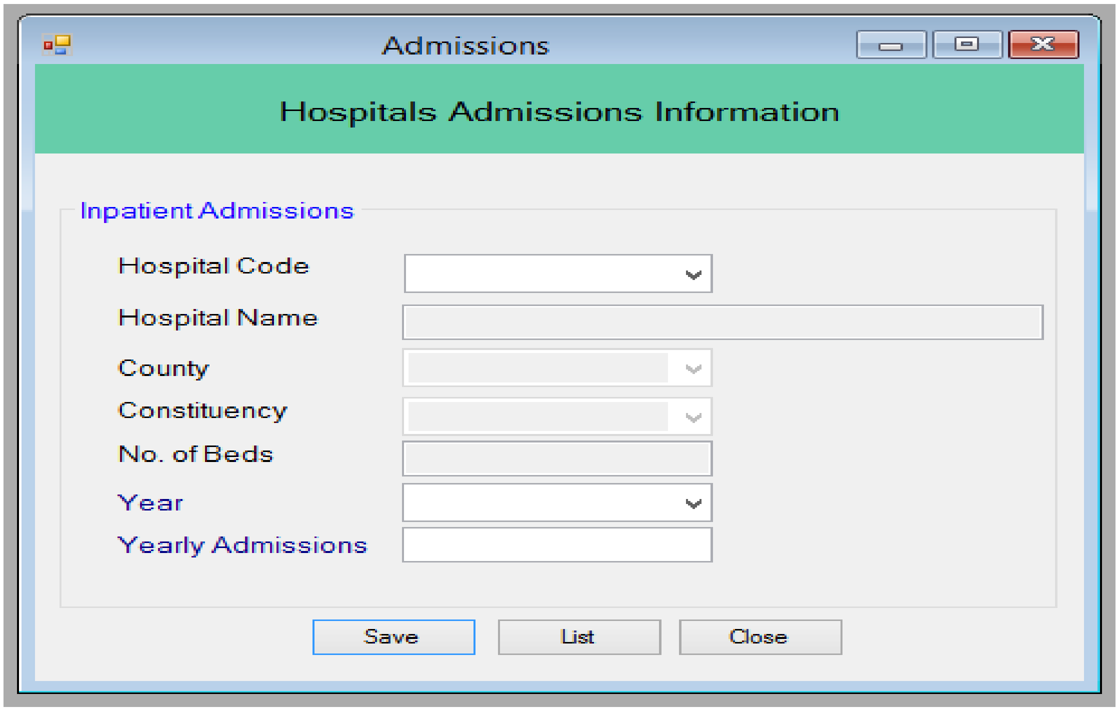


Figure 3.9: Sample Inpatients data

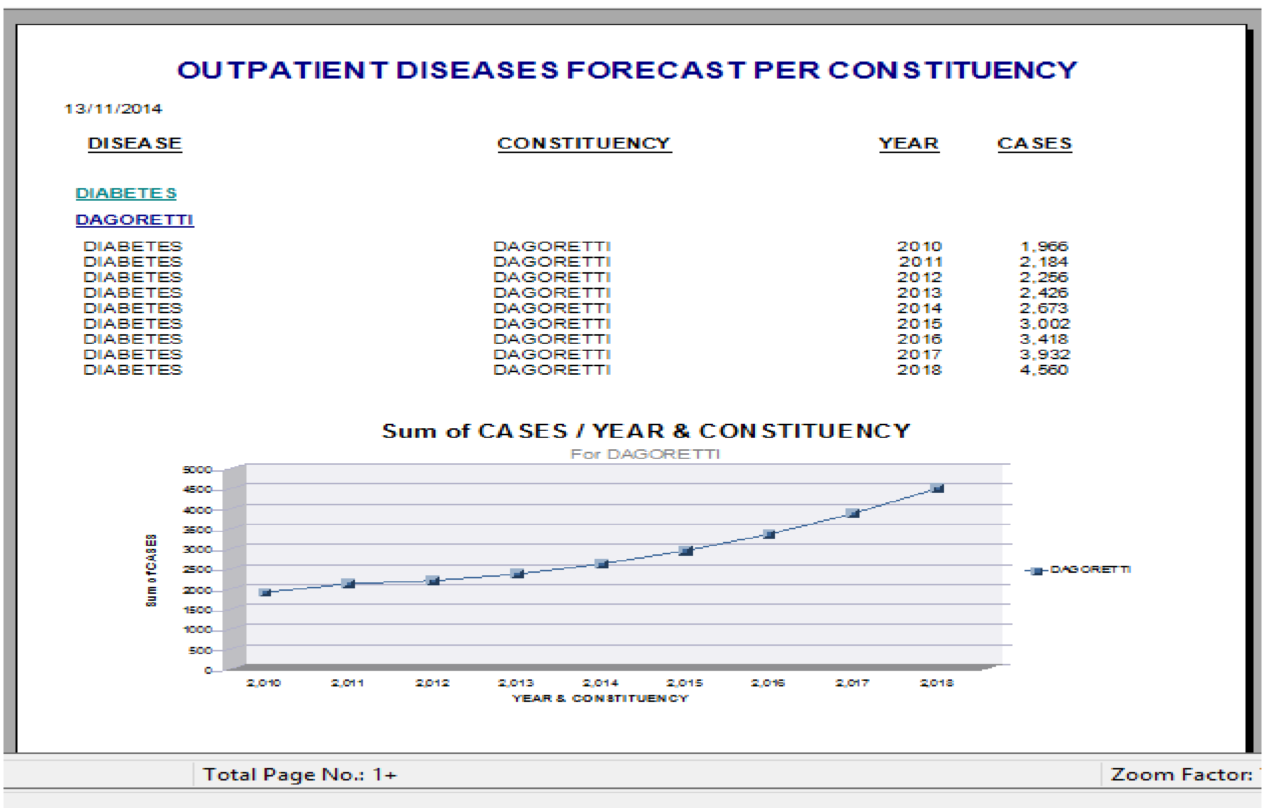


Figure 3.10: Sample Outpatients screen shot

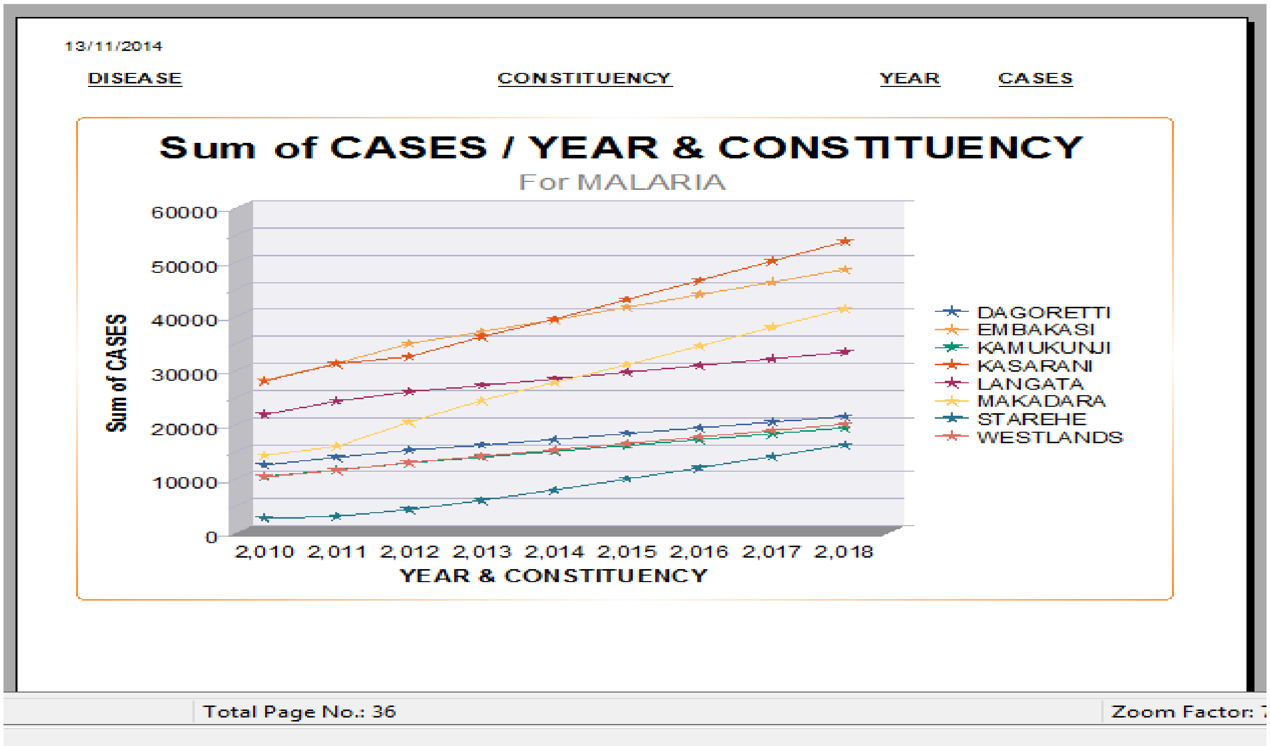


Figure 3.11 Outpatients Forecasts

3.9.4 Key algorithms

This solution implemented a few algorithms to realize the desired functionalities. These algorithms include the following:

1. WEKA Add on package for Time Series data Forecasts
2. Importing of Forecasts data into the model from WEKA
3. Value add to decision making using WEKA forecast and Decision Support system Parameters
4. Decision Support report generation

3.9.5 Code Model

Because of the complexity of solution, not all code will be presented in this report. For sample code, see Appendix A.

CHAPTER 4: EVALUATION

The data mining and knowledge discovery process in this research is aimed at forecasting the health services demand in Nairobi County for the purposes of fore planning by the health managers and planners. This chapter concentrates on evaluation and discussions of the model results. The results from the model are evaluated and validated. They are also compared to results from other demand forecasting methodologies in use in the health care sector.

4.1 System Evaluation

The evaluation of the system involved testing of the model that was implemented and the results from WEKA:

1. Unit testing
2. System testing

Unit Testing

This level of testing aimed at verifying the functionalities of different subsystems in connecting to database and reading and manipulating data

The tasks in Unit Testing:

Table 4.1: Tasks in Unit Testing

Task SNo.	Description
1	Running WEKA and Forecasting Package
2	Connection SQL Server Database Server from the model and WEKA through ODBC
3	Reading Data into WEKA from the database
4	Performing a Forecast
6	Running of Decision Support Reports

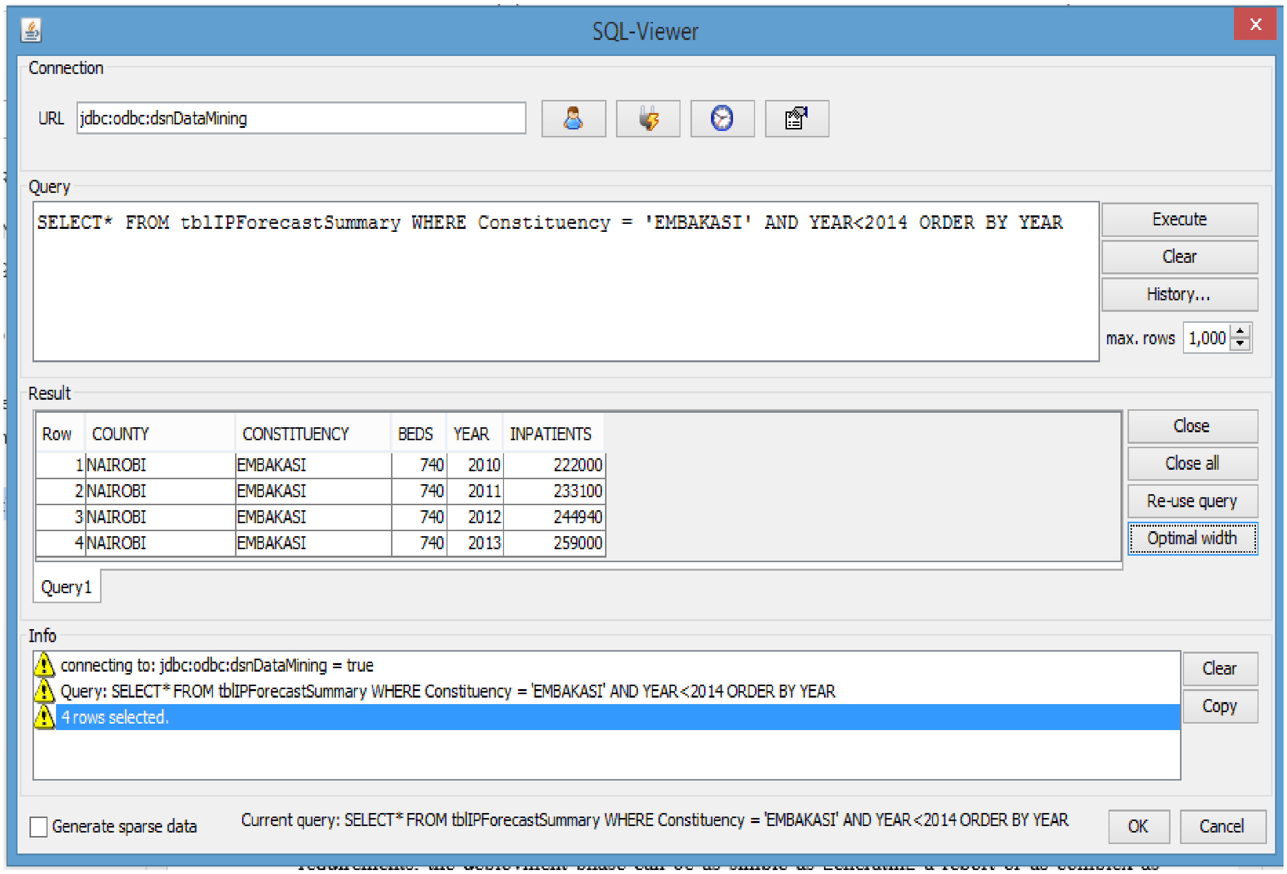


Figure 4.1: WEKA Database Connection and Query Execution

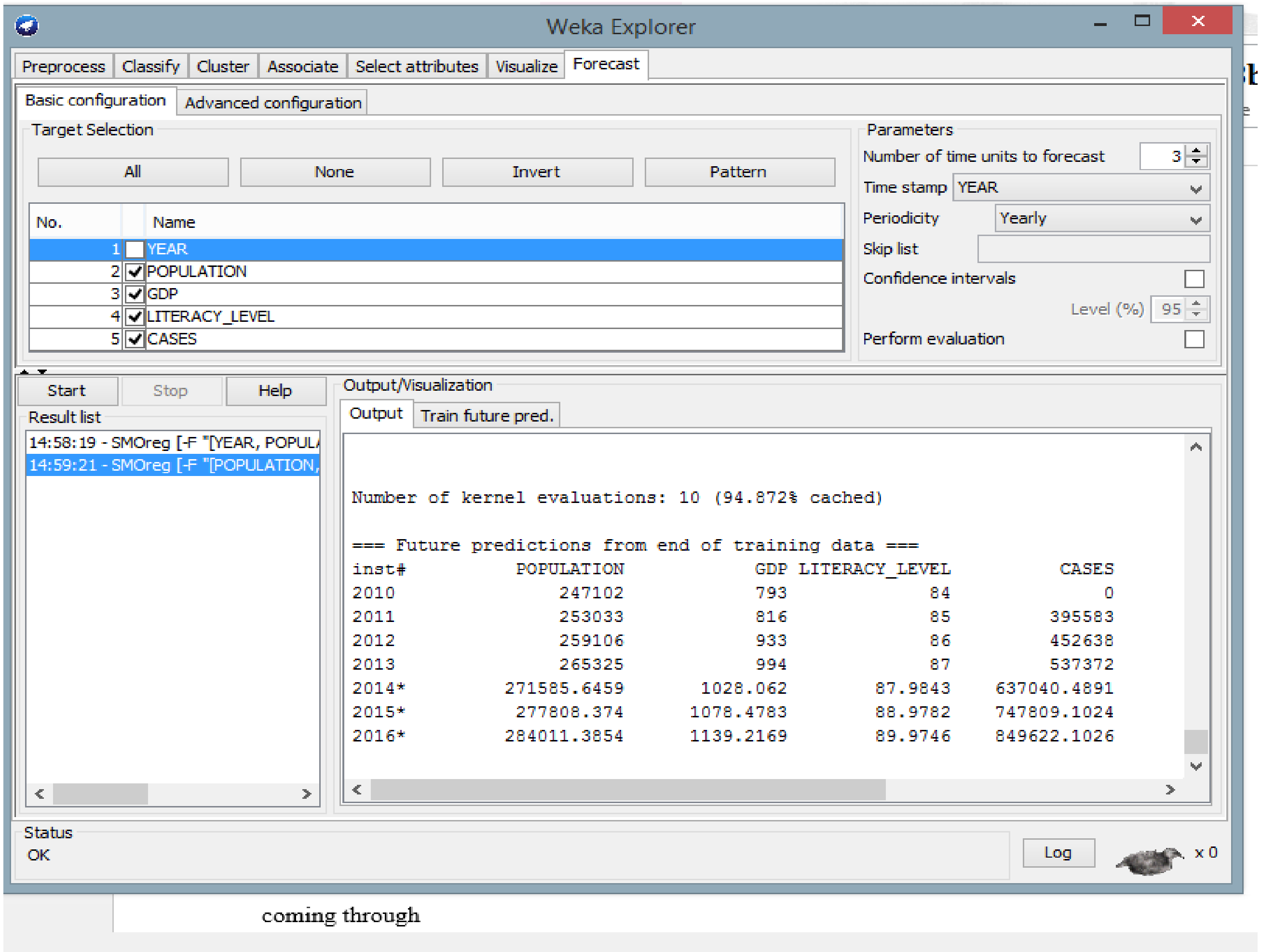


Figure 4.2: WEKA Results explorer

System testing

System testing was done to confirm that the system was doing the right functions and giving verifiable results. The forecasts were testing with the data to ascertain that the correct output was coming through

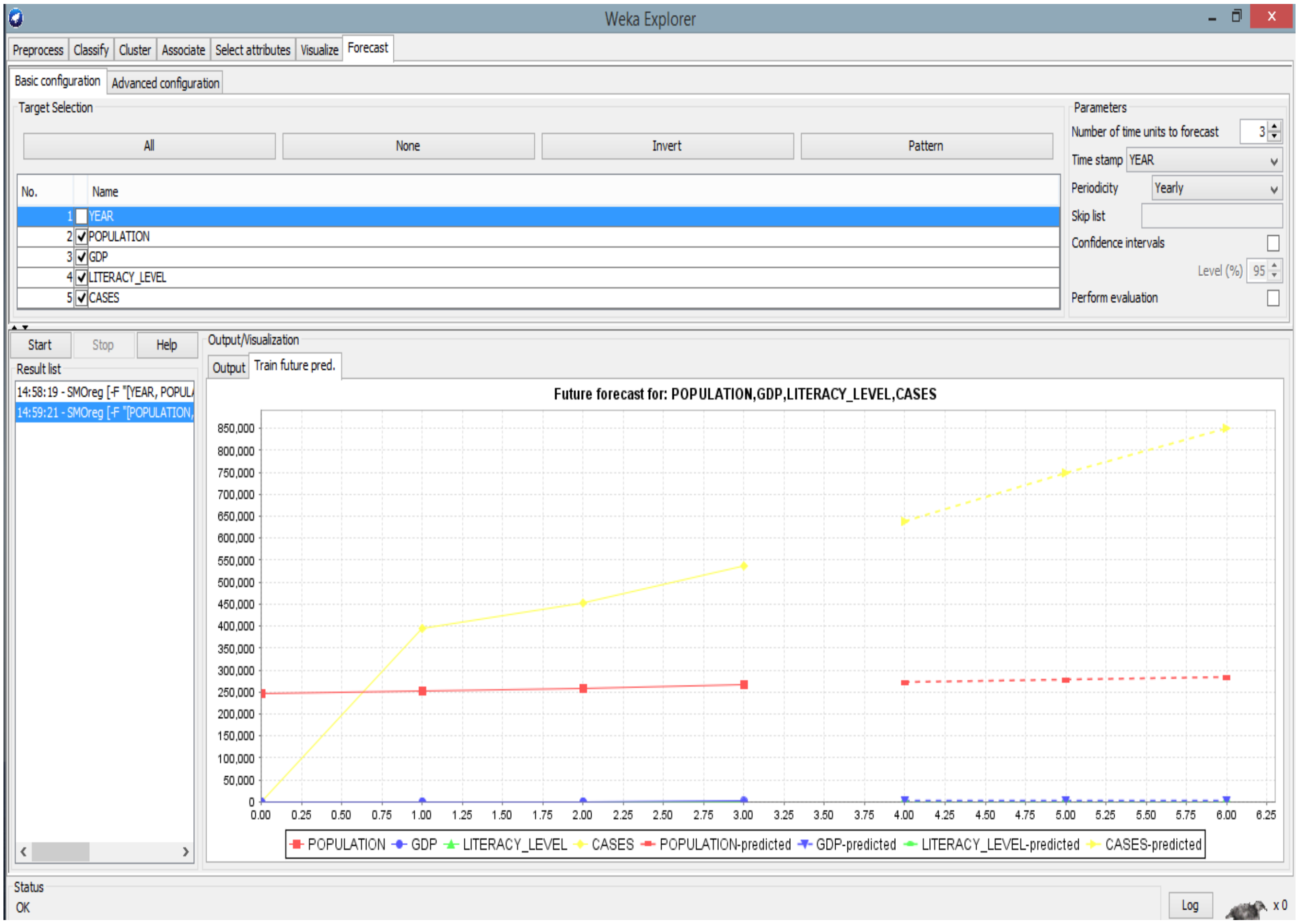


Figure 4.3: WEKA Forecast Graphs

4.2 Deployment

This solution will be deployed and tested in any private hospital in Nairobi County that would want to use demand forecasting functionality and decision support mechanism in this project. Depending on what the objective of such hospital, the deployment phase shall range from a simple as generation of a report to a complex implementation of a repeatable data mining process and real time integration to data sources in advanced servers and network infrastructure.

However, for this report, the prototype was deployed on a single laptop for demonstration purposes and design of the final project report

4.3 Demand Forecasting Results and Discussions

In this section, we describe and discuss the results that were obtained from the developed prototype. The purpose of this was to establish if the prototype met the functional requirements of the system and if the results can be relied on to make a decision. The results are grouped in three categories;

The Forecasting Package installed in WEKA supported multiple variables in the forecasting. This greatly improved the dependability of the results as opposed to regular linear forecasting methodologies used in other tools.

4.3.1 Total Outpatients Services Demand Forecasting Results

In outpatient total cases forecasting, we looked at the broad view of the totality of cases reported in each constituency and the number of admissions done from year 2011 to 2013. This data aided the forecasting of the inpatients in the next 3 years.

Table 4.2: WEKA 3 Years OP cases forecast

YEAR	COUNTY	CONSTITUENCY	POPULATION	GDP	LITERACY_LEVEL	CASES
2010	NAIROBI	DAGORETTI	329577	793	84	0
2010	NAIROBI	EMBAKASI	925775	793	84	0
2010	NAIROBI	KAMUKUNJI	261855	793	84	0
2010	NAIROBI	KASARANI	525624	793	84	0
2010	NAIROBI	LANGATA	355188	793	84	0
2010	NAIROBI	MAKADARA	218641	793	84	0
2010	NAIROBI	STAREHE	274607	793	84	0
2010	NAIROBI	WESTLANDS	247102	793	84	0
2011	NAIROBI	DAGORETTI	337487	816	85	479855
2011	NAIROBI	EMBAKASI	947994	816	85	527037
2011	NAIROBI	KAMUKUNJI	268140	816	85	342201
2011	NAIROBI	KASARANI	538239	816	85	309563
2011	NAIROBI	LANGATA	363713	816	85	453659
2011	NAIROBI	MAKADARA	223889	816	85	611267
2011	NAIROBI	STAREHE	281198	816	85	239239
2011	NAIROBI	WESTLANDS	253033	816	85	395583
2012	NAIROBI	DAGORETTI	345587	933	86	541725
2012	NAIROBI	EMBAKASI	970746	933	86	838213
2012	NAIROBI	KAMUKUNJI	274576	933	86	417561
2012	NAIROBI	KASARANI	551157	933	86	325396
2012	NAIROBI	LANGATA	372443	933	86	685836
2012	NAIROBI	MAKADARA	229263	933	86	755910
2012	NAIROBI	STAREHE	287947	933	86	310037

2012	NAIROBI	WESTLANDS	259106	933	86	452638
2013	NAIROBI	DAGORETTI	353882	994	87	524445
2013	NAIROBI	EMBAKASI	994044	994	87	1046417
2013	NAIROBI	KAMUKUNJI	281166	994	87	490559
2013	NAIROBI	KASARANI	564385	994	87	328389
2013	NAIROBI	LANGATA	381382	994	87	801659
2013	NAIROBI	MAKADARA	234766	994	87	790787
2013	NAIROBI	STAREHE	294858	994	87	397666
2013	NAIROBI	WESTLANDS	265325	994	87	537372
2014	NAIROBI	DAGORETTI	353882	994	88	**608849**
2014	NAIROBI	EMBAKASI	994044	994	88	**1268672**
2014	NAIROBI	KAMUKUNJI	281166	994	88	**563953**
2014	NAIROBI	KASARANI	564385	994	88	**334045**
2014	NAIROBI	LANGATA	381382	994	88	**936340**
2014	NAIROBI	MAKADARA	234766	994	88	**847972**
2014	NAIROBI	STAREHE	294858	994	88	**484697**
2014	NAIROBI	WESTLANDS	265325	994	88	**637040**
2015	NAIROBI	DAGORETTI	353882	994	89	**810942**
2015	NAIROBI	EMBAKASI	994044	994	89	**1489304**
2015	NAIROBI	KAMUKUNJI	281166	994	89	**637392**
2015	NAIROBI	KASARANI	564385	994	89	**339164**
2015	NAIROBI	LANGATA	381382	994	89	**1068127**
2015	NAIROBI	MAKADARA	234766	994	89	**900715**
2015	NAIROBI	STAREHE	294858	994	89	**571931**
2015	NAIROBI	WESTLANDS	265325	994	89	**747809**
2016	NAIROBI	DAGORETTI	353882	994	89	**1002299**
2016	NAIROBI	EMBAKASI	994044	994	89	**1710209**
2016	NAIROBI	KAMUKUNJI	281166	994	89	**710869**
2016	NAIROBI	KASARANI	564385	994	89	**344390**
2016	NAIROBI	LANGATA	381382	994	89	**1200381**
2016	NAIROBI	MAKADARA	234766	994	89	**954306**
2016	NAIROBI	STAREHE	294858	994	89	**659220**
2016	NAIROBI	WESTLANDS	265325	994	89	**849622**

Factors affecting the forecast results

1. Income – If the GDP remains constant, the results are different from if the capacity grows slightly. This showed that the forecasting mechanism integrated multiple variables in the forecast results.
2. Population and population growth – The population variations and growth affected the demand for the total outpatients seen in the forecast results.

3. Average Literacy Levels - This affected the results as it determines the average willingness to go and get treated in the private hospitals.

If the forecasting is done without including the values of these factors, the forecast seems to take a linear forecasting model which does not give the accurate forecast results as this only depends on time factor.

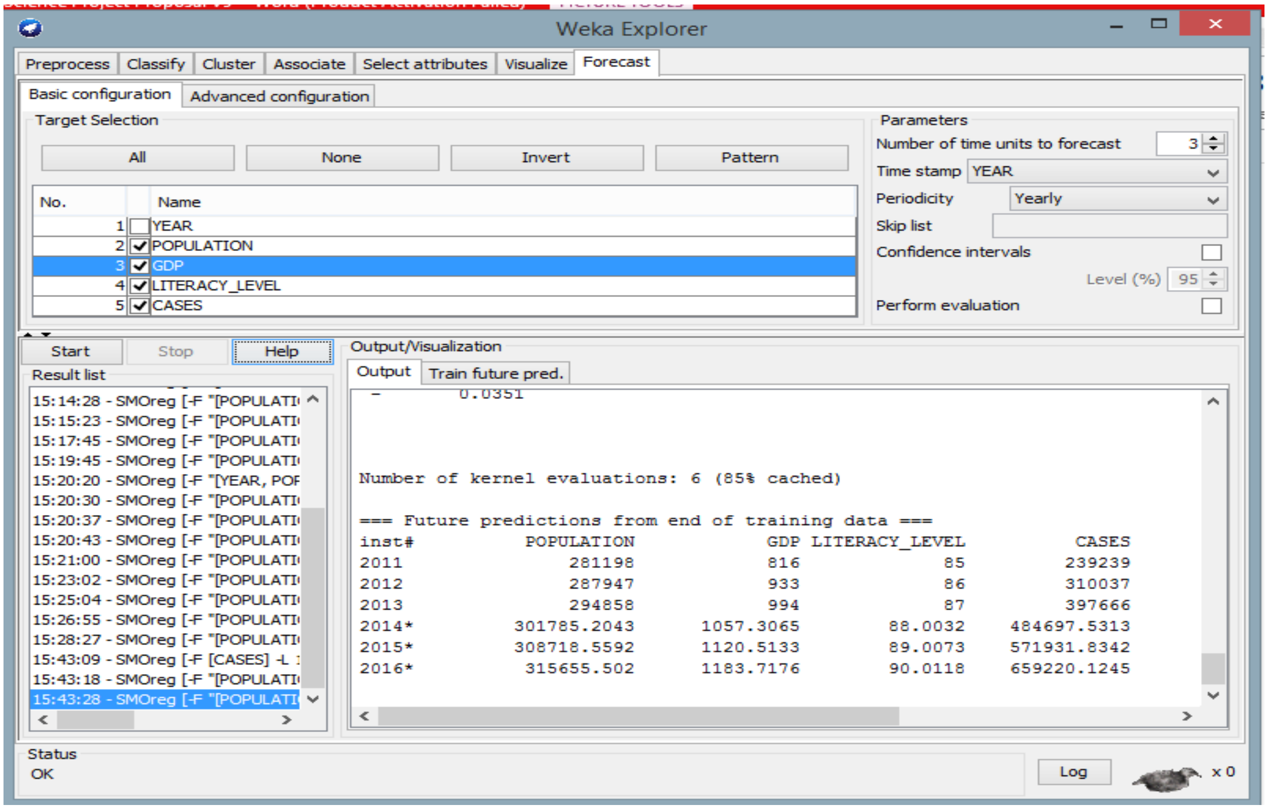


Figure 4.4: OP Results with all demand factors variations

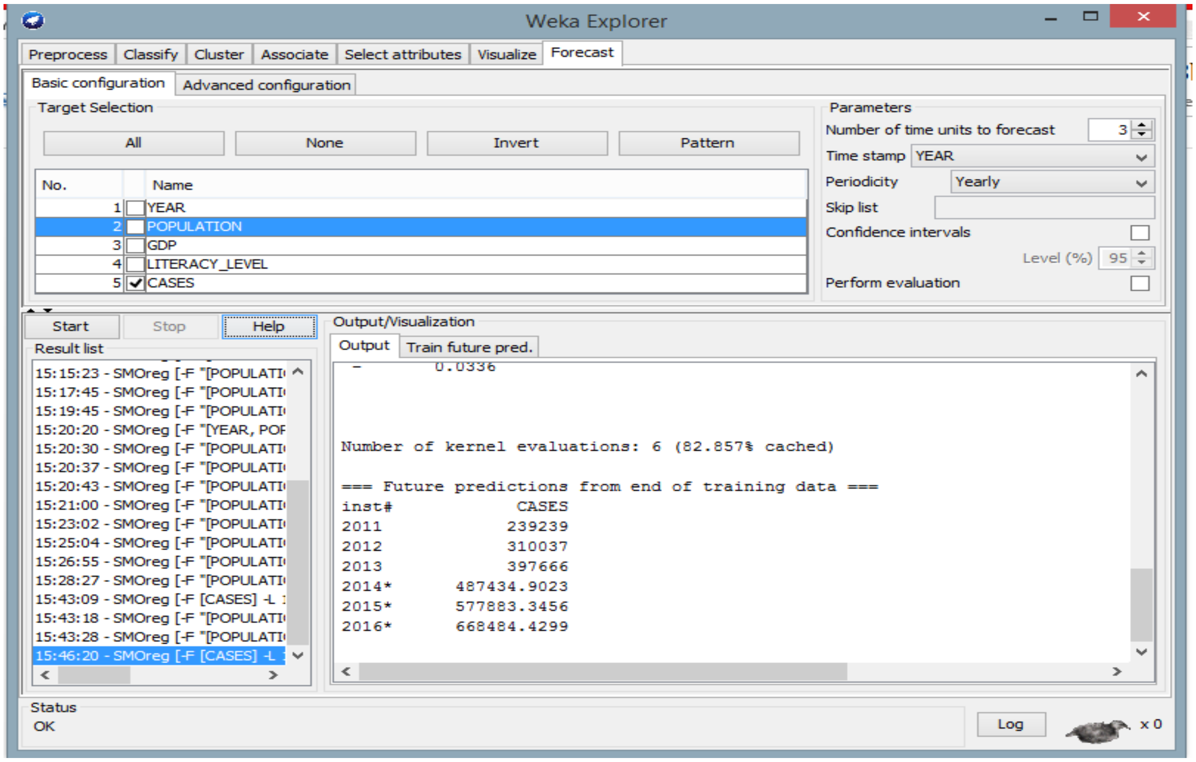


Figure 4.5: Results without Demand Factors Variations

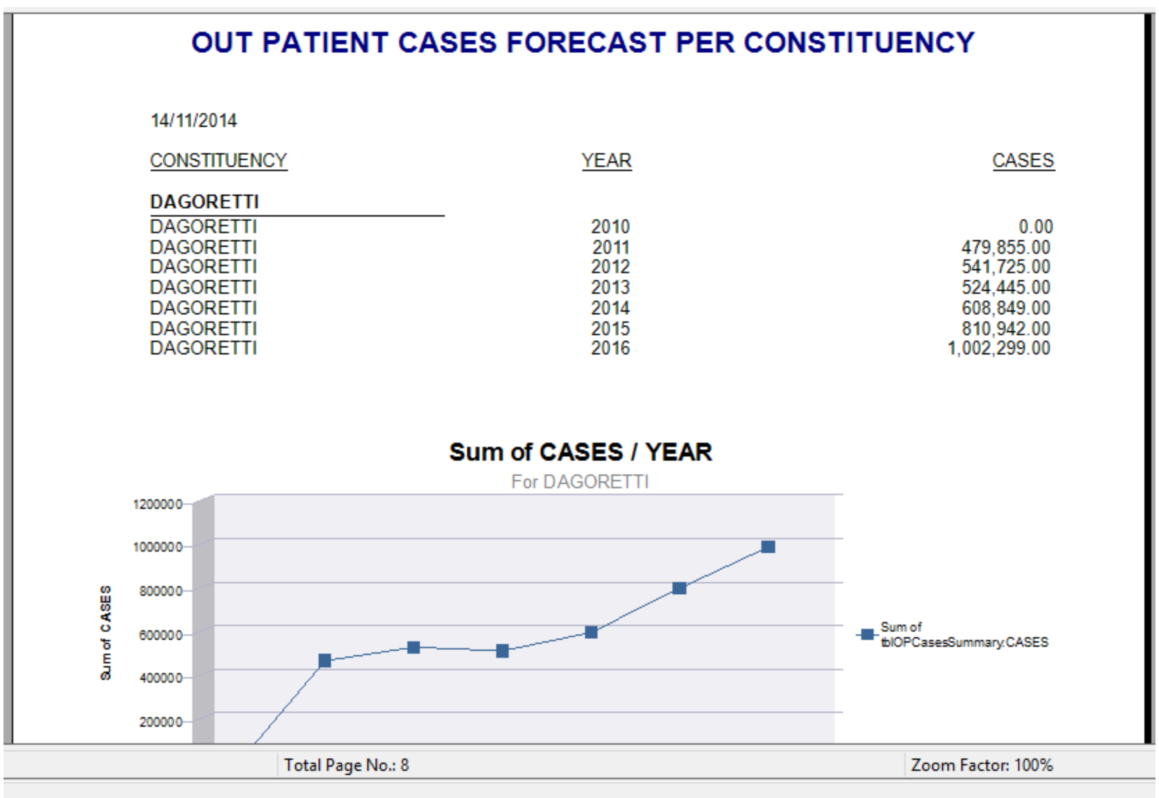


Figure 4.6: OP Forecast per constituency graph

4.3.2 Outpatient Specific Services Demand Forecasting Results

In Outpatient specific forecasting, we looked at both the broad view of the outpatient numbers in each constituency from year 2011 to 2013 as well as the specific cases that were treated. This data aided the forecasting of the inpatients in the next 3 years. Then we narrowed down to a few selected cases that are of interest to help the decision maker make a varied decision on what to do about the cases reported. These are;

1. Malaria
2. Diabetes
3. Hypertension

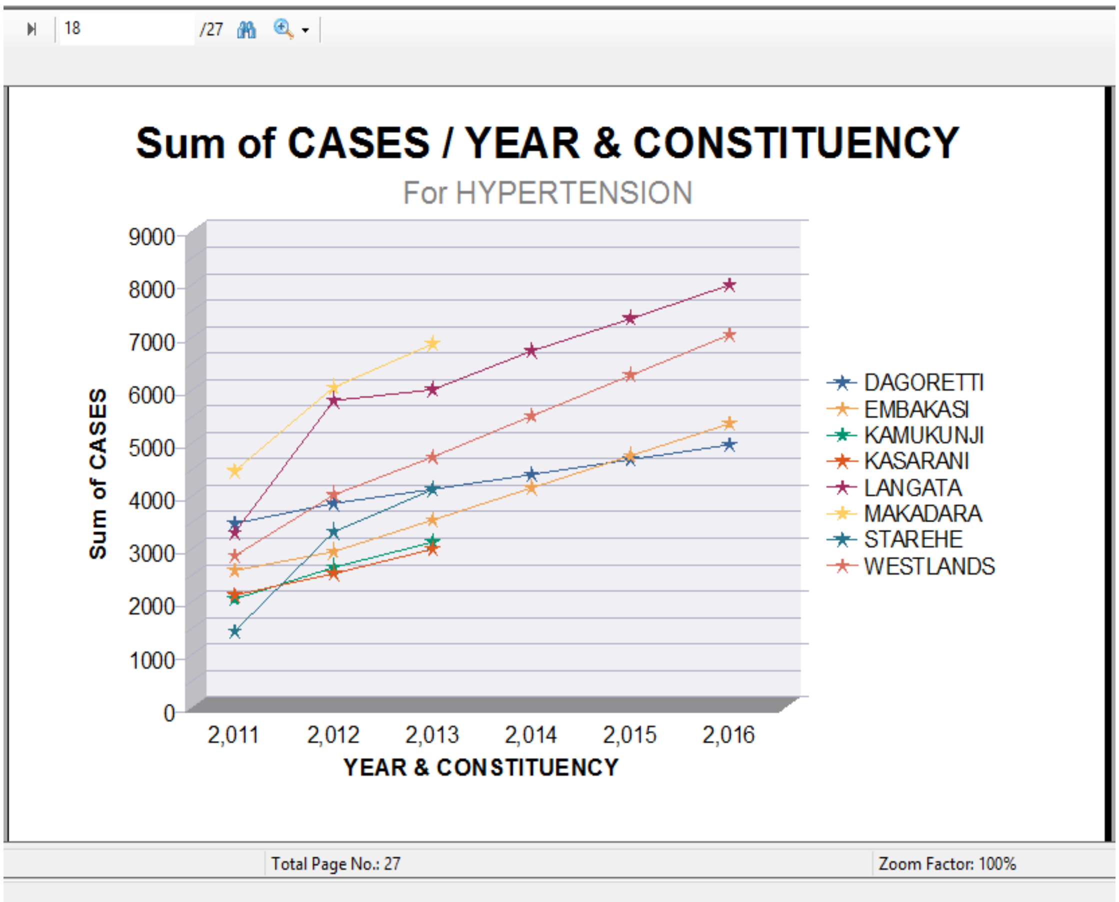


Figure 4.7: OP Summary Forecast

Factors affecting the forecast results

1. Income – If the GDP remains constant, the results are different from if the capacity grows slightly. This showed that the forecasting mechanism integrated multiple variables in the forecast results.

2. Population and population growth – The population variations and growth affected the demand for the total outpatients seen in the forecast results.
3. Average Literacy Levels - This affected the results as it determines the average willingness to go and get treated in the private hospitals.

4.3.3 Surgeries Demand Forecasting Results

In Surgeries forecasting, I looked at both the broad view of the surgeries numbers in each constituency from year 2010 to 2013 as well as the specific cases that were operated. I narrowed down to the following selected cases to assist the decision maker discern the trends and make the right investment decision;

1. Cardiac Open Heart Surgeries
2. Cancer Surgeries
3. Orthopedic Surgeries

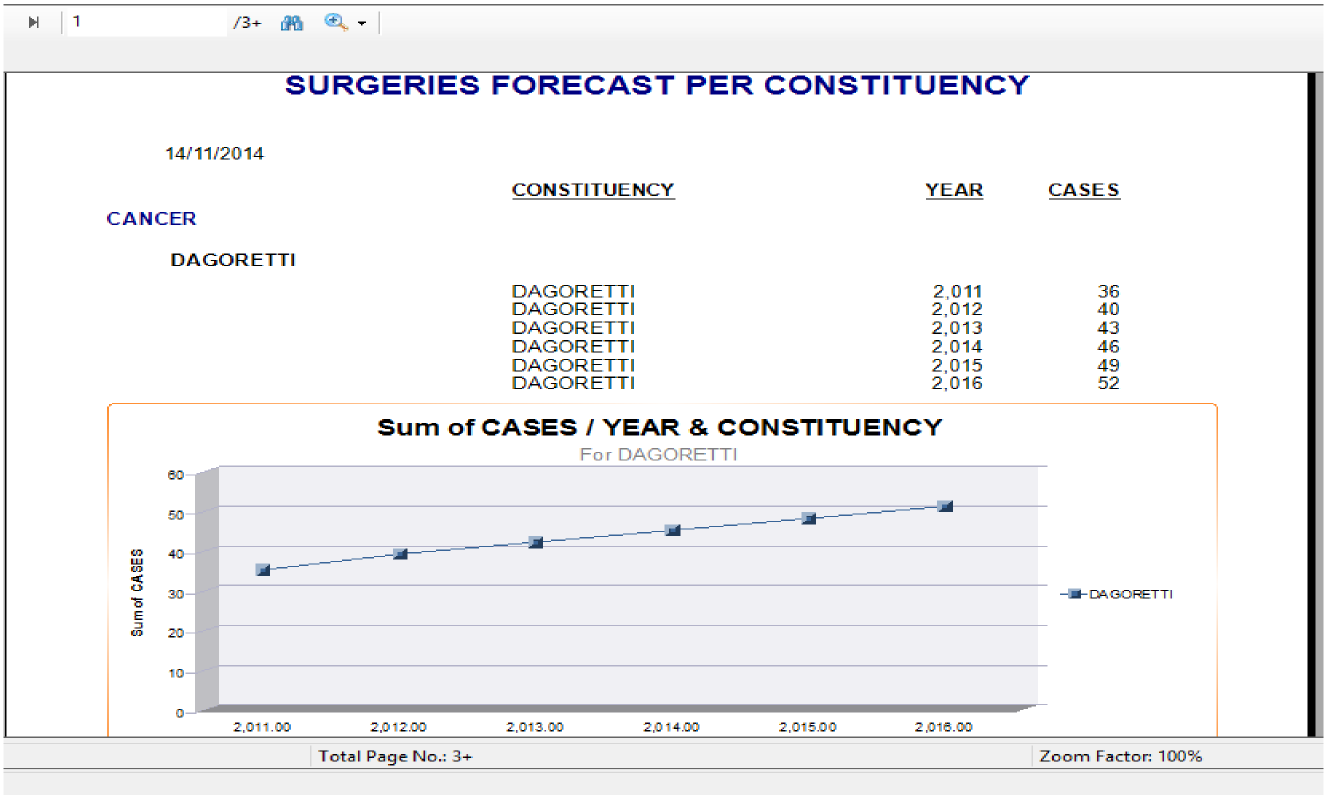
Table 4.3: Surgeries Forecast Results

YEAR	COUNTY	CONSTITUENCY	POPULATION	GDP	LITERACY_LEVEL	SURGERY	CASES
2011	NAIROBI	DAGORETTI	337487	816	85	HEART	22
2011	NAIROBI	DAGORETTI	337487	816	85	CANCER	36
2011	NAIROBI	DAGORETTI	337487	816	85	ORTHOPEDIC	147
2011	NAIROBI	LANGATA	363713	816	85	HEART	9
2011	NAIROBI	LANGATA	363713	816	85	CANCER	34
2011	NAIROBI	LANGATA	363713	816	85	ORTHOPEDIC	251
2011	NAIROBI	WESTLANDS	253033	816	85	HEART	16
2011	NAIROBI	WESTLANDS	253033	816	85	CANCER	30
2011	NAIROBI	WESTLANDS	253033	816	85	ORTHOPEDIC	123
2012	NAIROBI	DAGORETTI	345587	933	86	HEART	23
2012	NAIROBI	DAGORETTI	345587	933	86	CANCER	40
2012	NAIROBI	DAGORETTI	345587	933	86	ORTHOPEDIC	160
2012	NAIROBI	LANGATA	372443	933	86	HEART	25
2012	NAIROBI	LANGATA	372443	933	86	CANCER	59
2012	NAIROBI	LANGATA	372443	933	86	ORTHOPEDIC	268
2012	NAIROBI	WESTLANDS	259106	933	86	HEART	18
2012	NAIROBI	WESTLANDS	259106	933	86	CANCER	42
2012	NAIROBI	WESTLANDS	259106	933	86	ORTHOPEDIC	137
2013	NAIROBI	DAGORETTI	353882	994	87	HEART	25
2013	NAIROBI	DAGORETTI	353882	994	87	CANCER	43
2013	NAIROBI	DAGORETTI	353882	994	87	ORTHOPEDIC	169
2013	NAIROBI	LANGATA	381382	994	87	HEART	51
2013	NAIROBI	LANGATA	381382	994	87	CANCER	61

2013	NAIROBI	LANGATA	381382	994	87	ORTHOPEDIC	279
2013	NAIROBI	WESTLANDS	265325	994	87	HEART	48
2013	NAIROBI	WESTLANDS	265325	994	87	CANCER	49
2013	NAIROBI	WESTLANDS	265325	994	87	ORTHOPEDIC	149

FORECASTS

2014	NAIROBI	DAGORETTI	CANCER	**46**
2014	NAIROBI	DAGORETTI	HEART	**27**
2014	NAIROBI	DAGORETTI	ORTHOPEDIC	**178**
2014	NAIROBI	LANGATA	CANCER	**68**
2014	NAIROBI	LANGATA	HEART	**77**
2014	NAIROBI	LANGATA	ORTHOPEDIC	**290**
2014	NAIROBI	WESTLANDS	CANCER	**56**
2014	NAIROBI	WESTLANDS	HEART	**141**
2014	NAIROBI	WESTLANDS	ORTHOPEDIC	**161**
2015	NAIROBI	DAGORETTI	CANCER	**49**
2015	NAIROBI	DAGORETTI	HEART	**29**
2015	NAIROBI	DAGORETTI	ORTHOPEDIC	**188**
2015	NAIROBI	LANGATA	CANCER	**74**
2015	NAIROBI	LANGATA	HEART	**103**
2015	NAIROBI	LANGATA	ORTHOPEDIC	**304**
2015	NAIROBI	WESTLANDS	CANCER	**64**
2015	NAIROBI	WESTLANDS	HEART	**379**
2015	NAIROBI	WESTLANDS	ORTHOPEDIC	**173**
2016	NAIROBI	DAGORETTI	CANCER	**52**
2016	NAIROBI	DAGORETTI	HEART	**31**
2016	NAIROBI	DAGORETTI	ORTHOPEDIC	**197**
2016	NAIROBI	LANGATA	CANCER	**80**
2016	NAIROBI	LANGATA	HEART	**130**
2016	NAIROBI	LANGATA	ORTHOPEDIC	**314**
2016	NAIROBI	WESTLANDS	CANCER	**72**
2016	NAIROBI	WESTLANDS	HEART	**945**
2016	NAIROBI	WESTLANDS	ORTHOPEDIC	**185**


/3+
Total Page No.: 3+
Zoom Factor: 100%

SURGERIES FORECAST PER CONSTITUENCY

14/11/2014

CANCER

DAGORETTI

CONSTITUENCY	YEAR	CASES
DAGORETTI	2,011	36
DAGORETTI	2,012	40
DAGORETTI	2,013	43
DAGORETTI	2,014	46
DAGORETTI	2,015	49
DAGORETTI	2,016	52

Figure 4.8: Surgeries Forecast graph from model

Factors affecting the forecast results

1. Income – If the GDP remains constant, the results are different from if the capacity grows slightly. This showed that the forecasting mechanism integrated multiple variables in the forecast results.
2. Population and population growth – The population variations and growth affected the demand for the total outpatients seen in the forecast results.
3. Average Literacy Levels - This affected the results as it determines the average willingness to go and get treated in the private hospitals.

4.4 Testing the accuracy of the Prototype Forecast Results

For the forecast results to be dependable, they must be seen to be accurate and realistic. To test this on this model, the following procedures were done;

4.4.1 Performing the forecast for available data

Since from data collection and simulation I had data from 2011 to 2013, I set out to forecast the results for 2013 for which the results were know. The model forecasted the real results with a **7.9%** marginal error. See the table below.

Table 4.4: Forecast Results Margin Error

YEAR	CONSTITUENCY	OP CASES ACTUAL	WEKA OP FORECAST	WEKA % ERROR
2013	DAGORETTI	524445	563440	7.435479412
2013	EMBAKASI	1046417	1078609	3.07640262
2013	KAMUKUNJI	490559	502732	2.48145483
2013	KASARANI	328389	341189	3.897816309
2013	LANGATA	801659	891215	11.17133345
2013	MAKADARA	790787	830190	4.982757683
2013	STAREHE	397666	416699	4.786177345
2013	WESTLANDS	537372	509550	-5.177418995
	MEAN PERCENTAGE ERROR (MPE)			4.081750331

4.4.2 Performing comparison with other forecasting methods

The results were compared with output from Time series forecasts using ZAITUN Time Series tool that is an open source tool to perform linear forecast with Time Series data.

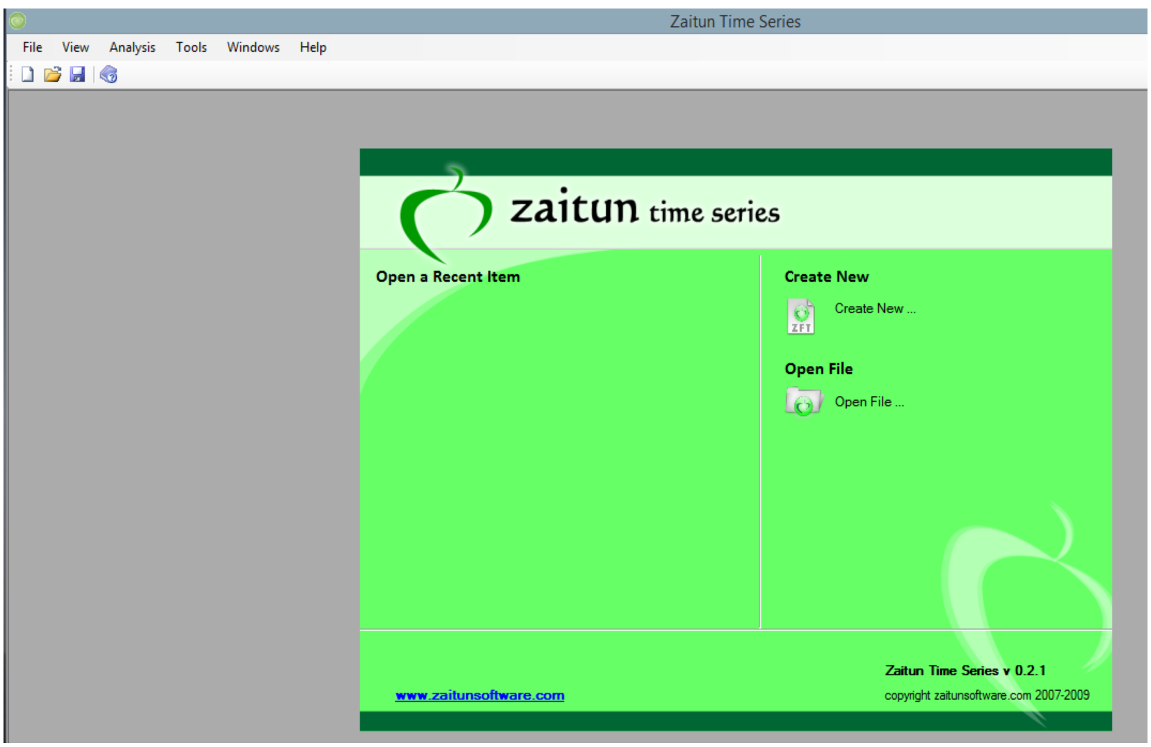


Figure 4.9: Zaitun Time Series Software Interface

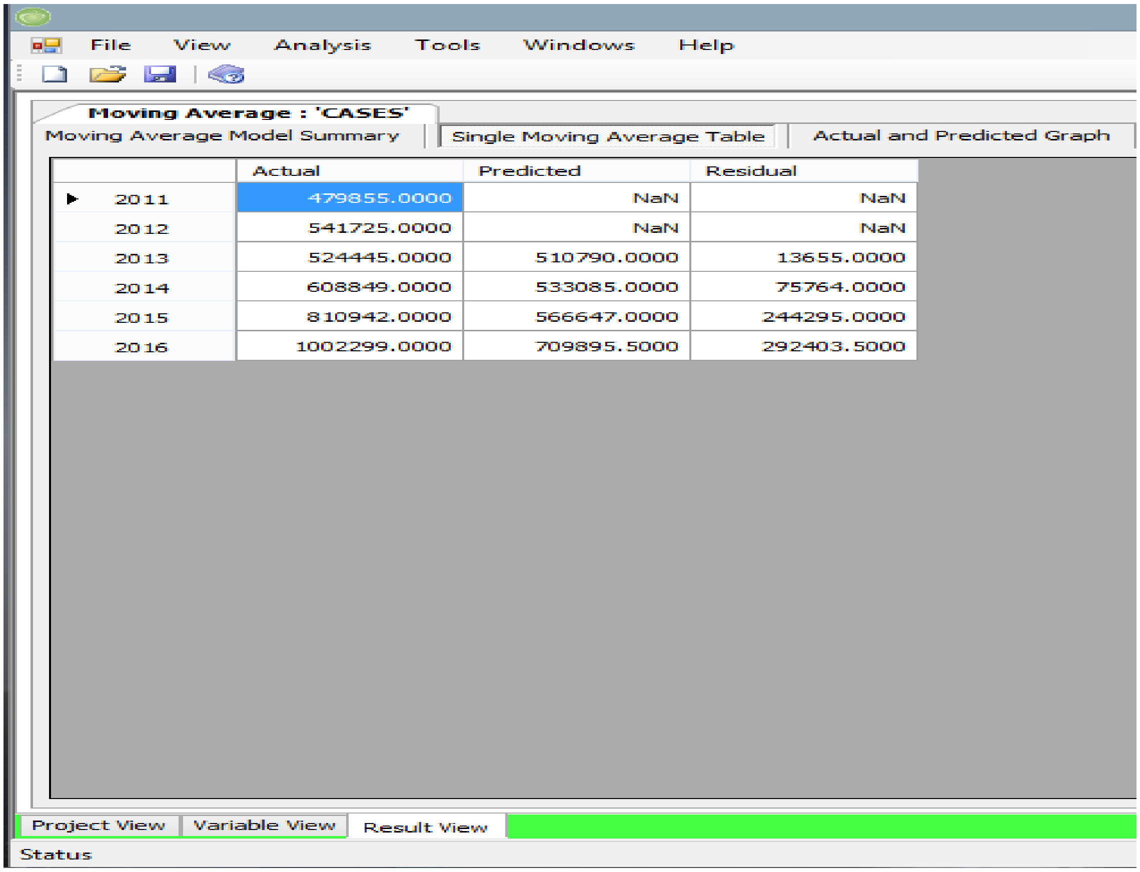


Figure 4.10: Zaitun Forecasts

After Analysis of same data on outpatients, the following were the results in comparison with the WEKA Neural Network Results;

Table 4.5: Margin Error for WEKAN, Moving Averages and Linear Regression

YEAR	CONSTITUENCY	OP CASES ACTUAL	WEKA OP FORECAST	WEKA % ERROR	MOVING AVERAGE OP FORECAST	MA % ERROR	LINEAR REGRESSION OP FORECAST	LR % ERROR
2013	DAGORETTI	524445	563440	7.435479412	510790	-2.6037049	475650.909	-9.3039482
2013	EMBAKASI	1046417	1078609	3.07640262	682625	-34.76549	1026629.82	-1.8909461
2013	KAMUKUNJI	490559	502732	2.48145483	379881	-22.561608	483871.091	-1.3633241

2013	KASARANI	328389	341189	3.897816309	317479.5	-3.3221271	318011.546	-3.1601103
2013	LANGATA	801659	891215	11.17133345	569747.5	-28.928946	790162.091	-1.4341396
2013	MAKADARA	790787	830190	4.982757683	683588.5	-13.555926	776722.727	-1.7785159
2013	STAREHE	397666	416699	4.786177345	274638	-30.93752	389693.727	-2.004766
2013	WESTLANDS	537372	509550	-5.177418995	274638	-48.892387	389693.727	-27.481572
	MEAN PERCENTAGE ERROR (MPE)			**4.081750331**		**-23.195964**		**-6.0521653**

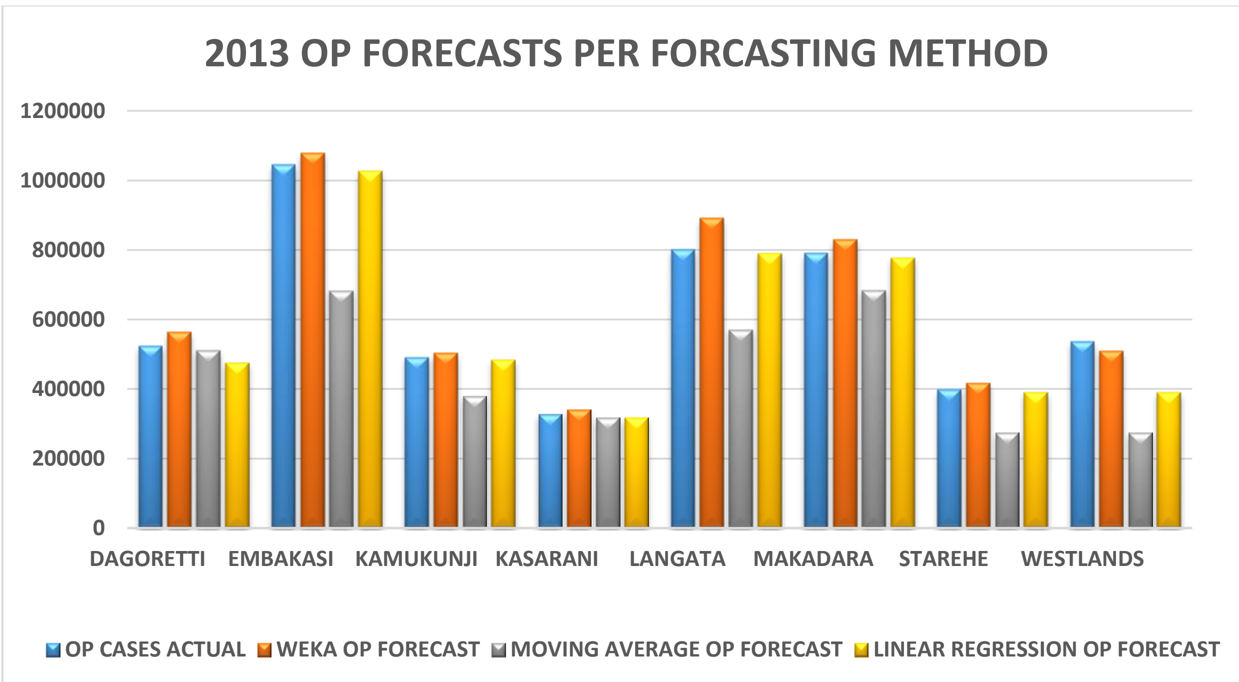


Figure 4.11: 2013 OP Forecasts per method

These results shows that WEKA Forecasts using Artificial Neural Network gives a more accurate forecast results that the Moving Averages and Linear Regression and hence gives a more reliable results for the demand forecast.

4.4.3 Testing if the system considered multiple variables

Testing if the system considered multiple variables – Since several factors affect the demand of the health services in Private hospitals as outlined above, this was also put into test. The results showed that if these factors were left out, the results were different from when included, and indeed lesser values of prediction

Table 4.6: Variables Elimination Forecasts

YEAR	CONSTITUENCY	DISEASE	CASES	WITHOUT POPULATION	WITHOUT GDP	WITHOUT LITERACY	WITHOUT ALL FACTORS
2011	EMBAKASI	DIABETES	1153				
2012	EMBAKASI	DIABETES	3169				
2013	EMBAKASI	DIABETES	5113				
2014	EMBAKASI	DIABETES	**7062**	**7061.1309**	**7053.8526**	**7060.786**	**7046.209**
2015	EMBAKASI	DIABETES	**9015**	**9013.5309**	**8999.8072**	**9012.9176**	**8987.7619**
2016	EMBAKASI	DIABETES	**10969**	**10966.8893**	**10946.6214**	**10965.9941**	**10930.0824**

These tests were used to confirm the accuracy of the forecast results from the model. The results of the model shows a trend where the patients visiting the private medical facilities is increasing steadily in all forecasts done.

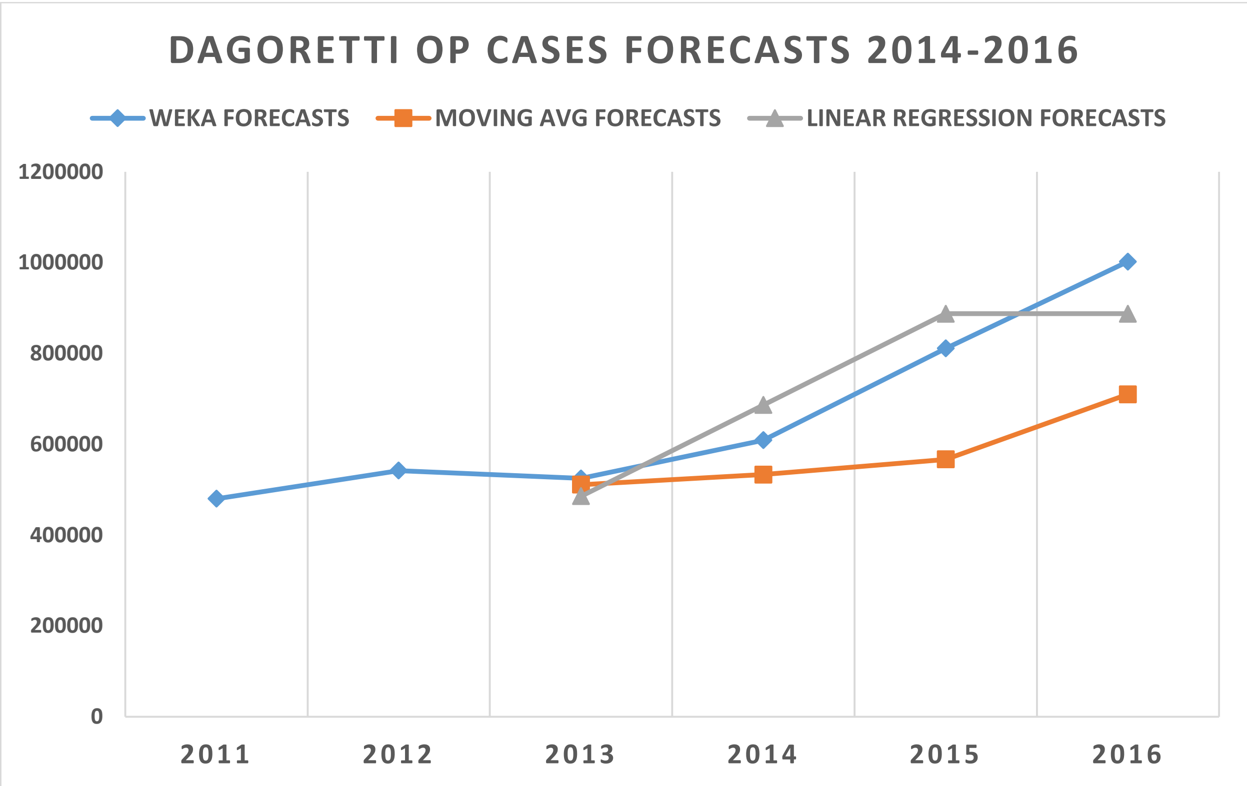


Figure 4.12: Dagoretti OP Cases Forecasts 2014-2016

CHAPTER 5: CONCLUSIONS & RECOMMENDATIONS

5.1 Conclusions

In Kenya today, the healthcare managers and planners must make future decisions about healthcare services delivery without knowing what will happen in the future. This because there does not exist an intelligent and sophisticated system that can mine this big data and provide analytical patterns on healthcare services demand forecast in Kenya. Forecasts would enable the managers to anticipate the future demand and plan accordingly.

The big and complex data in private medical facilities in Nairobi County has not adequately been used by the facilities for analysis and forecasting of demand and hence aiding of decision making process. Most of the hospitals in the county do not use any forecasting mechanism to assist fore planning of their services but open the hospitals for the sick to come for their services. For those who use forecasting, they use linear mechanism mainly based in Excel and does not reflect the many variables that affect the demand for their services by the clientele. They only look at the numbers and project future numbers which rarely happens to be realistic forecast and hence end up not using them for any decision making. They have no system that can aid in forecasting and help in decision making on future projections of demand.

The main objective of this project was to examine data mining as an approach to private health services demand forecast in Nairobi County. Specifically, we set out to explore the various forecasting methods in healthcare services demand, identify the most suitable healthcare services demand forecasting methods for Nairobi County and finally develop and validate a data mining model for forecast of private healthcare services demand for Nairobi County.

The effective knowledge discovery techniques and tools of Data mining in the modern world are at the epicenter of building any intelligent analysis and forecasting models from the big data in various industries. Data mining and forecasting tools like WEKA employed in this research and prototype building have proved to be very accurate in forecasting from the big data available in the medical world. The tool is able to perform complex functions to forecast future trends applying multiple variables and weighing them accordingly to reach at the forecast figures. Apart from forecasting procedures, the tool has a load of other functionalities that were not in the scope of this research which includes clustering, classifications and associations that can assist in discovering new patterns from the medical big data and aid decision making.

The Forecasting Package installed in WEKA supported multiple variables in the forecasting. This greatly improved the dependability of the results as opposed to regular linear forecasting methodologies used in other tools. The Factors that affect demand of private health care services in Kenya were reviewed and were seen to affect the forecast results. These factors are Income, population and Literacy levels. If the forecasting is done without including the values of these factors, the forecast seemed to take a linear forecasting model whose margin error was larger than including all factors.

For the forecast results to be dependable, they must be seen to be accurate and realistic. To test this on this model, forecasting of actual know data was done. Since from data collection we had data from 2011 to 2013, we set out to forecast the results for 2013 for which the results were know. The model forecasted the real results with a 7.9% marginal error which is expected to reduce with increase in training data. Testing if the system considered multiple variables – Since several factors affect the demand of the health services in Private hospitals as outlined above, this was also put into test. The results showed that if these factors were left out, the results were different from when included.
These two tests were used to confirm the accuracy of the forecast results from the model. The results of the model shows a trend where the patients visiting the private medical facilities is increasing steadily in all forecasts done.

A Data mining approach therefore to forecasting of demand for health services is the most accurate and ideal way of building a forecasting model that employs multiple variables. This was the main objective of this project - the use of data mining approach to build a healthcare services demand forecasting model in Nairobi County. The project proposed and implemented a prototype based on Data mining to forecast demand for health services in the private health facilities. The demand for services in the public facilities is affected by different factors to those that affect the demand in the private facilities. This lead to narrowing of research to only consider the private facilities since this is where all the factors will affect demand including purchasing power, population growth, investments and equipment’s put in place, time constraints and the class factor.

This prototype is therefore a very useful piece of invention that Private healthcare planners, investors and managers can use to forecast future demand and therefore plan their facilities accordingly in terms of equipment’s, new branches, staff training and management and cash flow.

5.2 Recommendations

The accurate functioning of this model largely depends on central availability of accurate data from all hospitals in the county for all medical cases attended to. My recommendations to the county government and the implementers of Kenya Health Information System in the Ministry of health is to get all data from all hospitals and in the country to enhance the functioning of forecasting models built from such data. This will aid any advancement in this model to be able to include more specific cases for a health Nairobi county. The sharing of data across all the interested parties should be enhanced so that data required to aid any such future works is made possible.

This project prototype has been built using VB.NET 2012 from Microsoft Visual Studio and SQL Server Database while WEKA is implemented using JAVA platform. This meant the forecasting classes in WEKA cannot be seamlessly accessed from the prototype and therefore, forecast data had to be uploaded in the database manually. This reduces the flexibility of scenarios that the use can try within the model outside of WEKA in case such data had not been uploaded. I recommend that this model in future be built in JAVA to aid a seamless integration of WEKA with the model. Or use JAVA bridges to execute the JAVA classes from Visual Studio.

5.3 Future Work

- This project confined the research to private hospitals in the county. The model can be expanded in future to include the public facilities and also facilities in all other 47 counties in Kenya.
- This model can also be enhanced in future so that a hospital can load its data and use it to perform a local forecast procedures of the trends in the hospital for decision making as well on matters expansion, staff training and development and equipment investments within its premises.
- GIS tools can be employed in future to visualize the forecasts in a GIS map.

REFERENCES

Abdulkader, A. (2006, May). Creating a GIS application for health services at Jeddah city. *Computers in Biology and Medicine*.

Abernethy, M. (2010). Data mining with WEKA, Part 1: Introduction and regression. Retrieved March 12, 2014, from http://www.ibm.com/developerworks/library/os-weka1

Balaji, D. (2011). Forecasting The Demand Of Pharmaceutical Product Based On Its Sales Record. *National Conference on Emerging Research and Advances in Mechanical Sciences* (pp. 13-18). Chennai: Dept. of Mechanical Engineering College of engineering, Guindy Anna University.

Chendroyaperumal, C. (2009, January). DEMAND FORECASTING TECHNIQUES: AN EVALUATION. *National Institute of Management Studies*. Retrieved March 8, 2014, from www.researchgate.net

Hans, L. (2013). *Predictive Analytics for Demand Forecasting and Planning Managers – A Big Data Challenge.* (S. K. KAIST College of Business, Ed.) Delphus Inc, .

Harleen, K., & Siri, K. W. (2006). Empirical Study on Applications of Data Mining Techniques in Healthcare. *Journal of Computer Science, 2*(2), 194-200.

Hian, C. K., & Gerald, T. (2005). Data Mining Applications in Healthcare. *Journal of Healthcare Information Management, 19*(2). Retrieved February 10, 2014, from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.92.3184&rep=rep1&type=pdf

Hursh, G., & etal. (1994). Private Providers' Contributions to Public Health in Four African Countries. *Conference - Private and Nongovernment Providers: Partners for Public Health in Africa.* Nairobi.

Karanja, F., & Wanyoro, A. (2010). UNDERSTANDING THE SPATIAL PREVALENCE OF CERVICAL CANCER USING GIS IN NAIROBI, KENYA. *Department of Geospatial and Space Techcnology, University of Nairobi*.

Kesten, C., & Armstrong, J. (2012). *DEMAND FORECASTING: EVIDENCE-BASED METHODS* (165 ed.).

Mary, K. O. (2004). Application of Data Mining Techniques to Healthcare Data. *STATISTICS FOR HOSPITAL EPIDEMIOLOGY, 25*(8).

Meenu, D., & Priyanka, D. (2013, July). Applications of Data Mining Techniques: Empowering Quality Healthcare Services. *International Journal of Information, Communication and Computing Technology Jagan Institute of Management Studies, 1*(1).

Mirjana, B. P., & Dijana, C. (2008). Data mining usage in health care management: literature survey and decision tree application. *Medicinski Glasnik, 5*(1).

Moturi, C., & Kioko, F. (2013, February). Use of Artificial Neural Networks for Short-Term Electricity Load Forecasting of Kenya National Grid Power System. *International Journal of Computer Applications (0975 – 8887), 63*(2).

Mouhib, A., & Wael, A. (2008). Using Data Mining Techniques for Predicting Future Car market Demand. *Information and Communication Technologies: From Theory to Applications, 2008. ICTTA 2008. 3rd International Conference* (pp. 1-5). Damascus: IEEE. Retrieved May 2, 2014, from www.researchgate.net

Neelam, S. (2006). *Forecasting for Global Health: New Money, New Products & New Markets.*

Patrizia, C., & Karin, B. (2013). *The full view - Planning/Forecasting/S&OP technology report - June 2013.* Manufacturing & Logistics IT Magazine. Retrieved March 15, 2014, from http://www.logisticsit.com/articles/2014/11/11/the-full-view-planning/forecasting/s-and-op-technology-report-june-2013/

Prasanna, D., Kuo-Wei, H., & Jaideep, S. (2011). DATA MINING FOR HEALTHCARE MANAGEMENT. *2011 SIAM International Conference on Data Mining.*

Reden, & Anders. (2007). Key Components of Demand Forecasting Hospital System Perspective. *2007 Health Spring Meeting; Demand Forecasting for Health Care Delivery Systems, 84.*

Rudiger, W., & Jochen, H. (2000). CRISP-DM: Towards a Standard Process Model for Data Mining. *Proceedings of the Fourth International Conference on the Practical Application of Knowledge Discovery and Data Mining* (pp. 29-39). Ulm, Germany: DaimlerChrysler Research & Technology FT3/KL. Retrieved February 7, 2014, from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.198.5133&rep=rep1&type=pdf

Salim, D., & etal. (2013). Overview Applications of Data Mining In Health Care: The Case Study of Arusha Region. *International Journal of Computational Engineering Research, 3*(8).

Sarab, A., Marko, G., & Dietrich, W. (2003). ECML/PKDD-2003 Knowledge Discovery Standards.

Timothy, D. R., & Justin, K. (2013). Using Data Mining in Forecasting Problems. *SAS Global Forum 2013, 085-2013.*

Yasar, A. O. (2005). Quantitative Methods in Health Care Management Chapter 2. In *Quantitative Methods in Health Care Management* (pp. 10-48). John Wiley & Sons, 2005. Retrieved February 12, 2014, from http://labs.fme.aegean.gr/decision/images/stories/docs/healthcare_forecast.pdf

APPENDICES

Appendix A – Sample Module Code

```
Module modDataMining

    'Myconnection Declarations
    Public ConnDataMining As New System.Data.Odbc.OdbcConnection
    Public ConnNestedDataMining As New System.Data.Odbc.OdbcConnection
    Public ConnXNestedDataMining As New System.Data.Odbc.OdbcConnection
    Public ConnXXNestedDataMining As New System.Data.Odbc.OdbcConnection

    Public CmdDataMining As New System.Data.Odbc.OdbcCommand
    Public CmdNestedDataMining As New System.Data.Odbc.OdbcCommand
    Public CmdXNestedDataMining As New System.Data.Odbc.OdbcCommand
    Public CmdXXNestedDataMining As New System.Data.Odbc.OdbcCommand

    Public drDataMining As System.Data.Odbc.OdbcDataReader
    Public drNestedDataMining As System.Data.Odbc.OdbcDataReader
    Public drXNestedDataMining As System.Data.Odbc.OdbcDataReader
    Public drXXNestedDataMining As System.Data.Odbc.OdbcDataReader

    Public LinkString As String
    Public LinkNestedString As String
    Public LinkXNestedString As String
    Public LinkXXNestedString As String


    Public Sub MaConn()

        ConnDataMining.ConnectionString = "FIL=MS Access;DSN=dsnDataMining"
        ConnNestedDataMining.ConnectionString = "FIL=MS Access;DSN=dsnDataMining"
        ConnXNestedDataMining.ConnectionString = "FIL=MS Access;DSN=dsnDataMining"
        ConnXXNestedDataMining.ConnectionString = "FIL=MS Access;DSN=dsnDataMining"



        Try
            ConnDataMining.Open()
```

```
            ConnNestedDataMining.Open()
            ConnXNestedDataMining.Open()
            ConnXXNestedDataMining.Open()
        Catch Exp As Exception
            MsgBox(Exp.Message, MsgBoxStyle.Critical, "Database Connection Error")

        End Try

    End Sub
    Public Sub LinkTable()

        Try
            CmdDataMining = New System.Data.Odbc.OdbcCommand(LinkString, ConnDataMining)
            CmdDataMining.ExecuteNonQuery()
            CmdDataMining.Dispose()
        Catch Exp As Exception
            MsgBox(Exp.Message, MsgBoxStyle.Critical, "Database Connection Error")
        Finally
            CmdDataMining.Dispose()
        End Try
    End Sub
    Public Sub NestedLinkTable()

        Try
            CmdNestedDataMining   =   New   System.Data.Odbc.OdbcCommand(LinkNestedString,
ConnNestedDataMining)
            CmdNestedDataMining.ExecuteNonQuery()
        Catch Exp As Exception
            MsgBox(Exp.Message, MsgBoxStyle.Critical, "Database Connection Error")
        Finally
            CmdNestedDataMining.Dispose()
        End Try
    End Sub
    Public Sub NestedXLinkTable()

        Try
            CmdXNestedDataMining   =  New   System.Data.Odbc.OdbcCommand(LinkXNestedString,
ConnXNestedDataMining)
            CmdXNestedDataMining.ExecuteNonQuery()
```

```
        Catch Exp As Exception
            MsgBox(Exp.Message, MsgBoxStyle.Critical, "Database Connection Error")
        Finally
            CmdXNestedDataMining.Dispose()
        End Try
    End Sub

    Public Sub NestedXXLinkTable()

        Try
            CmdXXNestedDataMining = New System.Data.Odbc.OdbcCommand(LinkXXNestedString,
ConnXXNestedDataMining)
            CmdXXNestedDataMining.ExecuteNonQuery()
        Catch Exp As Exception
            MsgBox(Exp.Message, MsgBoxStyle.Critical, "Database Connection Error")
        Finally
            CmdXXNestedDataMining.Dispose()
        End Try

    End Sub
    Public Sub main()
        Dim MyMain As New frmMain

        MaConn()

        Application.Run(MyMain)
    End Sub

    Public Sub ReadTable()
        CmdDataMining = Nothing
        Try
            CmdDataMining = New System.Data.Odbc.OdbcCommand(LinkString, ConnDataMining)
            'CmdDataMining.ExecuteNonQuery()
            drDataMining = CmdDataMining.ExecuteReader()

        Catch Exp As Exception
            MsgBox(Exp.Message, MsgBoxStyle.Critical, "Database Connection Error")

        Finally
```

```
            'CmdDataMining.Dispose()
        End Try
    End Sub



    Public Sub NestedReadTable()

        Try
            CmdNestedDataMining    =    New    System.Data.Odbc.OdbcCommand(LinkNestedString,
ConnNestedDataMining)
            'CmdNestedDataMining.ExecuteNonQuery()
            drNestedDataMining = CmdNestedDataMining.ExecuteReader()

        Catch Exp As Exception
            MsgBox(Exp.Message, MsgBoxStyle.Critical, "Database Connection Error")

        Finally
            'CmdSQL.Dispose()
        End Try
    End Sub
    Public Sub NestedXReadTable()

        Try
            CmdXNestedDataMining    =   New   System.Data.Odbc.OdbcCommand(LinkXNestedString,
ConnXNestedDataMining)
            'CmdXNestedDataMining.ExecuteNonQuery()
            drXNestedDataMining = CmdXNestedDataMining.ExecuteReader()

        Catch Exp As Exception
            MsgBox(Exp.Message, MsgBoxStyle.Critical, "Database Connection Error")

        Finally
            'CmdSQL.Dispose()
        End Try
    End Sub

    Public Sub NumeralsOnly(ByVal e As System.Windows.Forms.KeyPressEventArgs)
        Dim c As Char
```

```
        c = e.KeyChar
        If Not (Char.IsDigit(c) Or Char.IsControl(c)) Then
            e.Handled = True
            Beep()
        End If
    End Sub
    Public Sub XNumeralsOnly(ByVal e As System.Windows.Forms.KeyPressEventArgs)
        Dim c As Char
        c = e.KeyChar
        If Not (Char.IsDigit(c) Or c = "." Or Char.IsControl(c)) Then
            e.Handled = True
            Beep()
        End If
    End Sub

    Public Sub LettersOnly(ByVal e As System.Windows.Forms.KeyPressEventArgs)
        Dim c As Char
        c = e.KeyChar
        If Not (Char.IsLetter(c) Or Char.IsWhiteSpace(c) Or Char.IsControl(c)) Then
            e.Handled = True
            Beep()
        End If
    End Sub

    Public Sub LettersAndDigitsOnly(ByVal e As System.Windows.Forms.KeyPressEventArgs)
        Dim c As Char
        c = e.KeyChar
        If   Not   (Char.IsLetter(c)   Or   Char.IsWhiteSpace(c)   Or   Char.IsDigit(c)   Or
Char.IsControl(c)) Then
            e.Handled = True
            Beep()
        End If
    End Sub

    Public Function CapitalizeFirstLetter(ByVal MyNameString As String) As String

        'simple declaration for the spliting of the string
        Dim wd As String
        Dim i As Integer
```

```vb
        Dim Str As String

        'initialize the string
        Dim path As String = MyNameString
        Dim delimiters() As Char = {CChar(" ")}
        Dim parts() As String
        parts = path.Split(delimiters)
        Dim iPart As IEnumerator
        iPart = parts.GetEnumerator

        ' Loop through the parts of the sub divided string
        Str = ""
        While iPart.MoveNext()
            'ignore white spaces and the empty string
            If Trim$(iPart.Current.ToString) <> "" And Trim$(iPart.Current.ToString) <> "
" Then
                'capitalize the first Letter
                wd = iPart.Current(0).ToString.ToUpper()
                'Loop as you Concartinate the rest of the letters of the word
                For i = 1 To Len(iPart.Current.ToString.ToLower()) - 1
                    wd = wd & iPart.Current(i).ToString.ToLower()
                Next
                Str = Str & wd & " "
            End If
        End While
        'Return out the capitalized string
        CapitalizeFirstLetter = Trim$(Str)

    End Function
```