IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Volume-5, Issue-11 E-ISSN: 2347-2693

Using Data mining for Forecasting Public Healthcare Services in India: a case study of Punjab

Parveen Singh[1*], Vibhakar Mansotra[2]

[1*] Department of Computer Science and IT, University of Jammu, Jammu, India
[2] Department of Computer Science and IT, University of Jammu, Jammu, India

**Corresponding Author: imparveen@yahoo.com*

Available online at: www.ijcseonline.org

Received: 19/Oct/2017, Revised: 30/Oct/2017, Accepted: 21/Nov/2017, Published: 30/Nov/2017

***Abstract*:** A big benefit of using data mining and knowledge management techniques is to create a dynamic knowledge rich health care environment. The application of Knowledge Discovery in Databases (KDD) can be done by skilled employees with good knowledge of health care industry. Thus, meaningful patterns and strategic solutions can be developed while working with massive quantities of data which can help to improve the quality of healthcare services offered to patients. This function is particularly useful for Insurance companies, Physicians, Pharmaceutical companies and by the Government health planners and management personals for the formulation of effective policies. However, there are a many issues that arise while dealing with such massive data, especially how this data can be analyzed in a reliable manner. The basic aim of Health Informatics is to take medical data from the real world and from all levels of human existence to help advance our understanding of health care facilities, medicine and medical practices. In this paper, we explored the Health care data of one of the Northern State of India, Punjab, available with HMIS database, using Big Data tools and approaches, which help in answering several critical questions with respect to healthcare facilities, for effective utilization and policy formulation of resources available. Data of Indoor Patient Department (IPD) and Outdoor Patient Department (OPD) from 2010 to 2017 has been used to forecast the number of patients in advance for coming years, taking into consideration most efficient model based on the accuracy of the forecasts, so that the planning is done well in advance for providing better health care facilities for the forthcoming patients.

***Keywords*:** Big Data, HMIS, Data mining, KDD, OPD, IPD, Time Series

I. INTRODUCTION

Data mining is an analytical technique that uses interface of database technology, pattern recognition, statistics, data visualization, machine learning and expert system. Healthcare generates huge amount of administrative data about health facilities viz about patients, hospitals, bed costs, claims, etc. This Huge amount of data generated by healthcare transactions is too complex and voluminous to be processed and analyzed by traditional methods. With the use of application of ICT, the public healthcare agencies have at their disposal huge volume of data, which on detail analysis can help in efficient decision making [1,2]. Thus, the challenge is to retrieve relevant information from this mountain of data and act upon it in a timely manner. This research paper focused upon the health care services provided to the patients in different public healthcare institutions of Punjab. The patients are mainly classified as Hospital Inpatient care (IPD) which requires that you have a medical problem that is serious enough for a doctor to admit you into the hospital for an overnight stay and OPD (Outpatient Follow-up Care) where patients are provided medical consultations and other allied services. To prepare government agencies in advance so as to provide better facilities, IPD and OPD data is used on which Time series analysis is applied to forecast the trend. A time series requires a database that consists of sequence of values or events those changes with time [4]. In this paper, we focus on the monthly IPD and OPD data of Punjab state from April-01-2010 to March-01-2017. A forecasting for the next twelve months has been carried out so that the public healthcare institutions could prepare in advance to provide high-quality facilities to the patients.

II. HEALTHCARE DELIVERY SYSTEM IN INDIA

India's Public Health System is a three-tier system namely Primary, Secondary and Tertiary level of health care. Primary health care is to provide preventive, curative and primitive services to the community [5]. Common and simple ailments are taken care of at that level only. Primary Health Centre and Sub-Centre function at this level. Secondary Health Care is to provide curative and specialized care to the community as well as works as First Referral Centre for PHC and Sub-Centre. Community Health Centre, Sub- Divisional Hospital and District-Hospital function as secondary level of health care. Tertiary Health Care is to provide super specialized as well as comprehensive services to the community for the

complex ailments. Medical college and apex center function as tertiary Centre. This setup of health care infrastructure is existing in the country.

The primary tier is designed to have three types of health care institutions, namely, a Sub- Centre (SC) for a population of 3000-5000, a Primary Health Centre (PHC) for 20000 to 30000 people and a Community Health Centre (CHC) as referral centre for every four PHCs covering a population of 80,000 to 1.2 lakh. The rural health care infrastructure has been developed to provide primary health care services through a network of integrated health and family welfare delivery system. The Ministry of Health and Family Welfare (MoHFW), Government of India has implemented the population norms for all the public health facilities under the NRHM are as under in Table 1[5].

Table 1: Population Norms implemented by MoHFW.

Health Institution	Population Norm (GOI)	
	Plain area	Hilly/Tribal area
Sub-Centre	5,000	3,000
PHC	30,000	20,000
CHC	1,20,000	80,000

III. PUBLIC HEALTHCARE IN PUNJAB

Punjab has 50,362 sq. km. geographical area and 2.77 Crore population (2011 census). There are 22 districts, 142 blocks and 15340 villages in Punjab state. The state has population density of 550 per sq. km. as against the national average of 312. The decadal growth rate of the state is 13.70% against 17.64 % for the country and the population of the state continues to increase at a much faster rate than the national rate [7]. The list of health institutions in the Punjab state are given in Table 2[5].

Table 2: List of Public Healthcare Institutions in Punjab [5]

SNo	INSTITUTION	NUMBER
1.	DISTRICT HOSPITALS	22
2.	SUB DIVISIONAL HOSPITALS	41
3.	BLOCK PHCs	118
4.	CHCs	151
5.	PHC's	446
6.	SUBSIDARY HEALTH CENTRES	1187
7.	URBAN DISPENSARIES	88
8.	REVAMPING CENTRES	39
9.	POLICE HOSPITALS	8
10.	ESI HOSPITALS/ESI DISPENSARIES	24
11.	DRME HOSPITALS	6
12.	MENTAL HOSPITAL	1
13.	OTHER HOSPITAL	1
	TOTAL	2097

The comparative facts of important health and demographic indicators of Punjab state vs. India are as follows in table 3[5].

Table 3: Health and Demographic Indicators of Punjab Vs. India.

S.No	Indicator	Punjab	India
1	Total Population (In Crore) (Census 2011)	2.77 Crore	1028.61
2	Decadal Growth (%) (Census 2011)	13.73	21.54
3	Crude Birth Rate (SRS 2013)	15.9	21.4
4	Crude Death Rate (SRS 2013)	7.2	7
5	Natural Growth Rate (SRS 2013)	9.0	14.4
6	Infant Mortality Rate (SRS 2013)	26	40
7	Maternal Mortality Rate (SRS 2010-12)	172	178
8	Total Fertility Rate (SRS 2012)	1.8	2.4

IV. DATA

This paper is mainly based on secondary data collected from NRHM portal from April-01-2010 to March-01-2017. The analysis carried out in this paper is based on the previous 7 years data to predict the data for the next 12 months, using Time Series analysis.

V. METHODOLOGY

In this work, we have applied Time series forecasting method, a learning algorithm to make prediction on a database to extract knowledge, to forecast the number of patients in advance for coming years, so that the planning is done well in advance for providing better health care facilities for the forthcoming patients. A time series is a set of evenly spaced evenly spaced, continuous, numerical data obtained at regular time periods. In the time series forecasting methods, the forecast is based only on past values and assumes that factors that influence the past and the present will continue influence the future. Time series methods of modelling believe that history repeats itself. By analysing the historical data, time series exhibit the characteristics like trends, seasonal and nonseasonal cycles, pulses and steps, outliers etc. The time series modelling techniques like exponential smoothing and autoregressive integrated moving average (ARIMA) are used for time series data for the purpose of prediction [6]. The main objective of time series analysis is to find out a pattern in the historical data or time series data and then extrapolate the pattern into the future; the forecast is based exclusively on past values of the variable. A time series is a series of observations on a variable measured at successive periods of time. The measurements may be taken every hour, day, week, month, or year, or at any other regular interval [6]. Exponential Smoothing model is one of the most accepted forecasting methods that are used to forecast the future time for a time series data that have no obvious trend or seasonality [3]. Exponential smoothing uses weighted values of previous series observations to estimate future values. As such, exponential smoothing is not based on a theoretical understanding of the data. It forecasts one point at a time,

adjusting its forecasts as new data comes in. The technique is useful for forecasting series that exhibit trend, seasonality, or both. There is a variety of exponential smoothing models that differ in their treatment of trend and seasonality [6]. Exponential Smoothing models are classified as either seasonal or nonseasonal. Seasonal models are only available if the periodicity defined using the Time Intervals node is seasonal. The seasonal periodicities are: cyclic periods, years, quarters, months, days per week, hours per day, minutes per day, and seconds per day. In Time Series Exponential Smoothing Criteria, the different model types are: Simple, Holt's Linear trend, Brown's Linear trend, Damped trend, Simple Seasonal, Winters' additive, Winters' multiplicative etc. In this paper, we choose Winters' additive model. Winters' additive model is suitable for a sequence in which there is a linear trend and a seasonal effect that is constant over time. Its relevant smoothing parameters are level, trend, and season. Winters' additive exponential smoothing is most similar to an ARIMA with zero orders of auto regression; one order of differencing; one order of seasonal differencing; and p+1 orders of moving average, where p is the number of periods in a seasonal interval. For monthly data, p=12[4,6].

VI. DATA ANALYSIS

The Data used for analysis is collected from HMIS portal from April-2010 to March-2018, for forecasting the future values. The IPD and OPD historical data set gives the trends and seasonality patterns that help us to decide the accurate model for predicting the future values and thus helps the IPD and OPD to make better decisions for the patients. The data analysis has been carried out by using IBM SPSS Modeler. The data stream to predict the patients for Outdoor Patient Department (OPD) for next four months is shown in Fig 1. The data file is in Excel format and the time interval is selected for 12 months for carrying out prediction of OPD patients services.

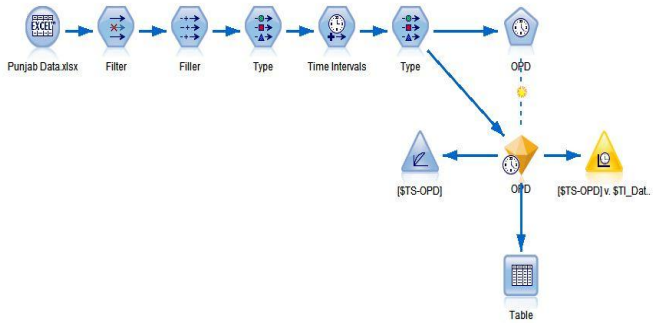

Fig. 1 Data Stream for predicting the patients for OPD for next 12 months.

The data stream to predict the patients for Indoor Patient Department (IPD) for next twelve months is shown in Fig. 2. The data file is in excel format and the time interval selected for prediction is 12 months. The statistical parameters are

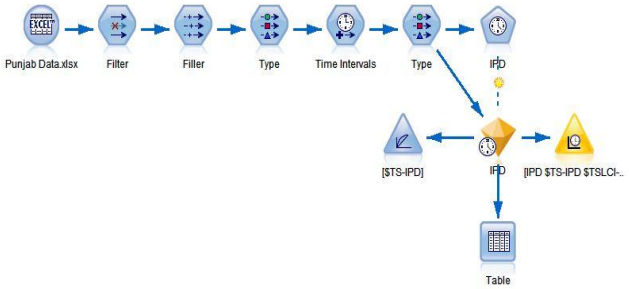

Fig. 2 Data Stream for predicting the patients for IPD for next 12 months.

being evaluated later on. The data has been taken from April 2010 to March 2018 and the actual and predicted values for OPD and IPD are shown with the help of time plot in figure 3 and figure 4 respectively. The time plot node allows viewing one or more time series plotted over time. The dots represent the historical data from April 2010 to March 2018 and the line represents the predicted values for the next twelve months.

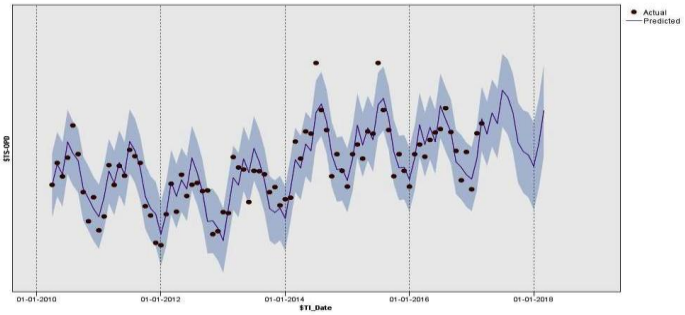

Fig. 3 Time Plot selected time series models for OPD.

The time plot shown above gives an indication that it is trending with the passage of time and also the number of OPD patients is also increasing.

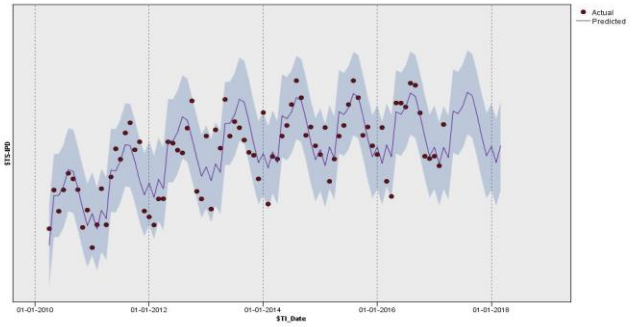

Fig. 4 Time Plot selected time series models for IPD.

The time plot shown in fig 4 provides a uniform pattern on account of seasonality which impacts the rise n fall of number of IPD patients.

Fig 5 represents the actual OPD values, predicted OPD values, Lower Confidence Intervals(LCI) OPD values and Upper Confidence Intervals(UCI) OPD values. The upper and lower limit provide a range of OPD patients that may increase with the passage of time depending upon the historical data. The graph takes into account both trends and seasonality following Winters Additive Exponential smoothening methods.

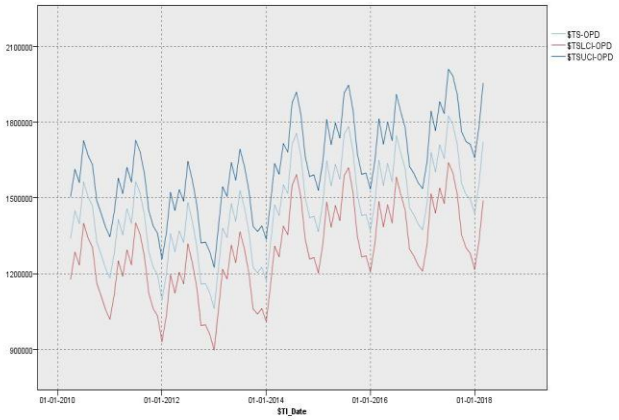

Fig. 5 Time Plot representing TS-OPD, LCL-OPD and UCL-OPD.

Figure 6 represents , Predicted IPD values, Lower Confidence Intervals(LCI) IPD values and Upper Confidence Intervals(UCL) IPD values using Time plot graph. The IPD prediction provide a glimpse of seasonality index in prediction. The trends shows that July and August month has high number of IPD patients on account of rainy season. Such seasonal indications provide a helpful tool to the healthcare planner to plan the resources and facilities accordingly.

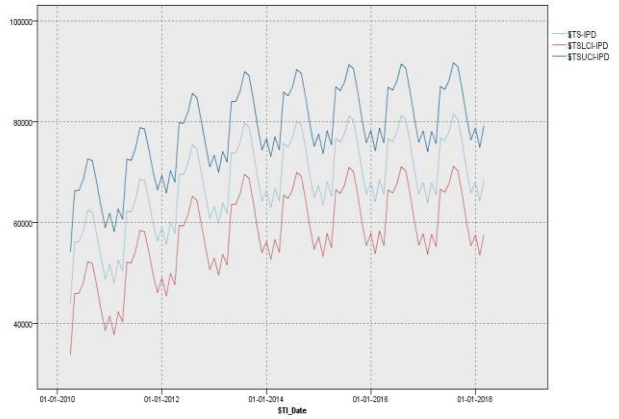

Fig. 6 Time Plot representing TS-IPD, LCL-IPD and UCL-IPD.

The Table 4 and Table 5 provide an account of predicted number of OPD and IPD patients for next complete year i.e. from Mar 2015 to Feb 2016. The upper and lower limit is also provided so that public healthcare institutions should provide minimal level of resources and facilities to the patients affected.

Table-4 Predicted Opd Values with Lower & Upper Limit (From April 2017 to Mar 2018)

Forecasting Period	Predicted OPD	LCIOPD	UCIOPD
Apr-2017	1602615.567	1439213.117	1766018.018
May-2017	1709278.114	1538450.082	1880106.146
Jun-2017	1655566.824	1477622.737	1833510.911
Jul-2017	1823906.735	1639120.357	2008693.114
Aug-2017	1790639.305	1599255.035	1982023.575
Sep-2017	1710287.267	1512525.041	1908049.493
Oct-2017	1557941.352	1354000.471	1761882.233
Nov-2017	1512839.57	1302901.737	1722777.403
Dec-2017	1495569.077	1279800.844	1711337.31
Jan-2018	1438168.599	1216723.362	1659613.836
Feb-2018	1558862.434	1331882.079	1785842.789
Mar-2018	1720941.601	1488557.876	1953325.326

Table-5 Predicted Ipd Values with Lower & Upper Limit (From April 2017 to Mar 2018)

Forecasting Period	Predicted IPD	LCI IPD	UCI IPD
Apr-2017	65431.355	55240.98	75621.73
May-2017	76811.511	66612.948	87010.075
Jun-2017	76232.869	66020.525	86445.213
Jul-2017	77971.291	67738.148	88204.434
Aug-2017	81443.924	71181.562	91706.286
Sep-2017	80547.059	70245.689	90848.429
Oct-2017	75913.272	65561.785	86264.759
Nov-2017	70431.849	60017.873	80845.826
Dec-2017	65895.217	55405.186	76385.248
Jan-2018	68142.52	57561.761	78723.279
Feb-2018	64230.899	53543.724	74918.073
Mar-2018	68294.213	57484.028	79104.397

Evaluation of Models: The forecasting models are evaluated on the basis of statistical parameters or goodness of measures. The Table 6 below shows a number of goodness-of-fit measures.

Table 6: Goodness –Of-Fit Measures

S. No	Target	Model	Stationary R**2	R**2	MAPE	Norm. BIC	Ljung-Box		
							Q	Df	Sig.
1	IPD	Winters additive	0.725	0.734	6.386	1,445.163	18.757	15.0	0.2
2	OPD	Winters additive	0.589	0.802	4.513	1,911.325	11.277	15.0	0.7

R**2 is the R-squared value, an estimation of the total variation in the time series that can be explained by the model. As the maximum value for this statistic is 1.0, both models are fine in this respect. The additional goodness-of-fit measure include the mean absolute percentage errors (MAPE). Absolute percentage error is a measure of how much a target series varies from its model-predicted level, expressed as a percentage value [12]. The MAPE value shows that all models display a mean uncertainty of less than 8%, which is very low. Interesting though these absolute values are, it is the values of the percentage errors (MAPE) which is more useful in this case, as the target series represent the healthcare services of various types. MAPE values represent an acceptable amount of uncertainty with the models. We have found that the goodness-of-fit statistics fall within acceptable bounds.

VII. CONCLUSIONS

In this paper, using time series data, we predict the number of expected patients in the next twelve months by using the historical data of previous year's w.e.f April 2010 to March 2018 with IBM SPSS Modeler. In addition to this, the lower interval and upper interval range of data is also predicted for the next twelve months from April 2017 to March 2018. Similarly, we can predict the future data for 2-5 years based on the historical data. Time series data mining is an integrated solution to forecast correct results that are totally based upon the accurate historical data. So time series data mining offers assurance in helping organizations to uncover hidden patterns in their data. In this paper, we study the behaviour of time series data using IBM SPSS Modeler and predicted the future data for the Indoor Patient Department and Outdoor Patient Department of Punjab state. We can predict the patients of IPD and OPD for more than 2-5 years and shall also predict the patients at national level. In addition to IPD and OPD, we can choose other indicators in healthcare to predict the future trends so that the healthcare planners could prepare in advance for providing various facilities to the patients.

REFERENCES

[1]. Rajesh Kumar Sinha, " *Impact of Health Information Technology in Public Health*", Sri Lanka Journal of Bio-Medical Informatics 2010,1(4):223-36

[2]. Pushpalata Pujari and Jyoti Bala Gupta, "*Exploiting Data Mining Techniques for Improving the Efficiency of Time Series data using SPSS-Clementine,*" Journal of Arts, Science & Commerce, vol. 3, issue 2(3), pp.69-80, April 2012.

[3]. The NRHM website. [Online]. Available: http://nrhm.gov.in/

[4]. http://censusindia.gov.in/vital_statistics/SRS_Bulletins/SRS%20Bulletin%20-Sepetember%202014.pdf

[5]. Albert Orwa Akuno et al., "*Statistical Models for Forecasting Tourists' Arrival in Kenya,*" Open Journal of Statistics, vol.5, pp.60-65, Feb. 2015.

[6]. Labib Arafeh, " *A Modified Neurofuzzy Based Quality of eLearning Model (Modified SCeLQM)*", International Journal of Computer and Information Technology , Vol 03 , No. 06, November 2014.

[7]. P. Singh and V. Mansotra, "*Data mining based Tools and Techniques in Public Health Care management: - A study*", in INDIACom2017, New Delhi, 2017, pp. 425- 29.

