

See discussions, stats, and author profiles for this publication at: <https://www.researchgate.net/publication/318316652>

# Handling Non Communicable Disease Using Predictive Analysis of Data Mining Techniques

Conference Paper · July 2017

CITATIONS

0

READS

63

2 authors:



Divya Zion

4 PUBLICATIONS 11 CITATIONS

SEE PROFILE



B.K. Tripathy

VIT University

736 PUBLICATIONS 4,269 CITATIONS

SEE PROFILE

Some of the authors of this publication are also working on these related projects:



clustering [View project](#)



Mental Health and Workload Management Platform [View project](#)

# Handling Non Communicable Disease Using Predictive Analysis of Data Mining Techniques

<sup>1</sup>Divya Zion, <sup>2</sup>B.K.Tripathy

School of Computer Science and Engineering, VIT University, Vellore-632014, Tamil Nadu, India

Email: [1ziondivya@gmail.com](mailto:1ziondivya@gmail.com), [2tripathybk@vit.ac.in](mailto:2tripathybk@vit.ac.in)

Received:9<sup>th</sup> June 2017, Accepted:25<sup>th</sup> June 2017, Published: 1<sup>st</sup> September 2017

## Abstract:

Non communicable diseases (NCDs), though not as dangerous as communicable diseases as far as proliferation is concerned, these are dangerous by themselves because of their root causes. In this article, our main focus is on predictive analysis of smoking, which is a non-communicable disease, using regression analysis. Regression analysis is a cornerstone of predictive based data mining techniques. We deploy Oracle Data Miner (ODM) which is a software mining tool for tracking smoking analysis report and Support Vector Machine (SVM) algorithm for experimental analysis using nonlinear regression. Risk factors for NCDs are obtained from Global Health Observatory (GHO) data of World Health Organization (WHO) and from reports on Tobacco control from National Tobacco Control Cell (NTCC) for detail survey. The datasets are carefully examined and different prevention methods or hazards to be faced thereof for different age groups are suggested. The treatments or precautions to be taken are dependent upon age groups. This study provides measures in reducing tobacco use and hence can be a key component in the concerned national or global plan.

**Key words:** Data mining, Predictive Regression Analysis, Oracle Data Miner Tool, SVM, National Tobacco Control Cell, Global Health Repository

## I. INTRODUCTION

Data mining provides versatile tools that automatically generate valuable information from data and transform that data into organized information. Today's challenging task in many organizations is to explore predictive analytics using statistics, data mining and game theory to analyze current and historical facts in order to make predictions about future events. There are various approaches to predictive analytics, and most of them depend on clean databases and the ability to mine data for patterns or to create classifications. This paper mainly focuses on risk factors and datasets of Non Communicable Diseases (NCD) [1] [4] [5] which are 'slow-moving catastrophe'. Predictive analysis [25] based data mining techniques are used to prevent the risk factors of NCD by using data obtained from GHO [2] [3] and the reports generated by tobacco control in India. Non Communicable Diseases are known as chronic diseases which are not passed

from person to person. These mining techniques are used to prevent the risk factors of NCD [4][5] by using data obtained from GHO and the reports generated by tobacco control in India. Non Communicable Diseases are known as chronic diseases which are not passed from person to person. These diseases are slow in progression and the main affecting areas in human body are cardiovascular i.e. heart attacks and stroke, chronic respiratory diseases (such as pulmonary disease and asthma), cancer and diabetes. Irrespective of the age group humans are affected with Non Communicable Diseases and are led to death.

The rise of Non Communicable Diseases in humans is propelled by some major risk factors [1] like; Tobacco use, Physical Inactivity, alcohol consumption, and taking unhealthy diet. The reasons for every risk factor are as follows:

## Reasons for Risk Factor of NCD

### 1) Use of Tobacco[7]

Smoking is the common process for consuming tobacco and tobacco is the substance mixed with additives and then combusted. The resulting smoke is inhaled and absorbed in the lungs. The substances that are present in cigarette smoke contain chemical reactions that result or yield in nerve endings, increases heart rate and reaction time, in lower age group it is less pronounced. The tobacco products are made entirely with leaf tobacco as raw material and contains highly addictive psychoactive ingredient which is nicotine. This tobacco usage leads to chronic diseases which is common throughout the world.

### 2) Physical Inactivity[8]

The physical inactivity is due to insufficient participation in physical activity during leisure time. The physical activity should not be confused with "exercise" which is a sub category of physical activity. Beyond any physical activity that is done during leisure time becomes as a part of persons work that has a health benefit. The adequate levels of physical activity improve functional health, reduce risk of hyper tension, heart diseases, and stroke and are fundamental to energy balance and weight control. All the above adequate levels of physical activity become inadequate when physical inactivity is insufficient.

### 3) Alcohol Consumption[10][11]

Alcohol is a depressant drug which slows down the vital functions. The cause of alcoholism is unknown and alcohol dependency develops when you drink so much and that chemical changes in the brain occur. There are certain factors

that may increase your risk for developing the diseases irrespective of the gender which include weight gain and high blood pressure from repeated over consumption of alcohol and raise to depression, liver damage, cancer, less immunity power.

**4) Unhealthy Diet[6][7]**

Poor eating habits include under-or-over eating, not having enough of healthy foods we need each day or taking too many types of food and drink which are low in fiber or high in fat, salt and sugar hence unhealthy diet is the major risk factors for chronic diseases like cardiovascular diseases, cancer, diabetes and problems linked to obesity and contributes to stress and high blood pressure. To have healthy diet specific recommendations should be taken like eating more fruits, vegetables, nuts and grains, cutting down salt, sugar and fats. Below given image gives an idea on how each risk factor leads to diseases in slow progression and affects the areas in human bodies.

AGE	NCD RISK FACTORS			
	TOBACCO USE	PHYSICAL INACTIVITY	ALCOHOL CONSUMPTION	UNHEALTHY DIET
15	20%	20%	10%	70%
20	70%	20%	30%	70%
25	98%	30%	90%	85%
30	98%	30%	80%	85%
35	80%	40%	70%	85%
40	40%	40%	70%	60%
45	40%	60%	60%	40%
50	30%	70%	40%	40%
55	30%	80%	20%	30%
60	30%	85%	20%	30%
65	30%	90%	20%	30%

Table:1 Percentage of NCD risk factors irrespective of the age

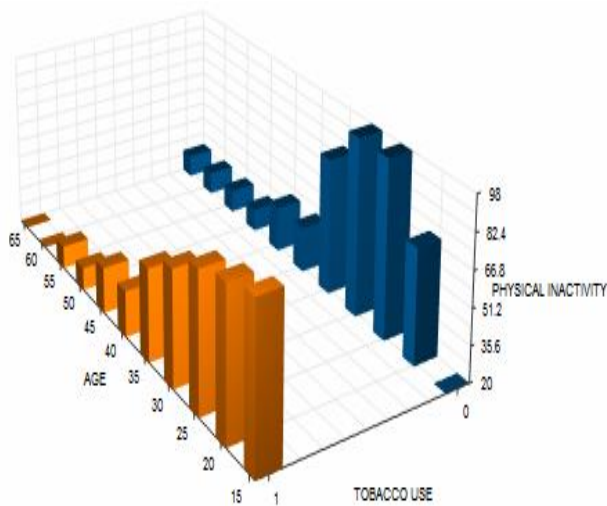


Fig.1 Graph which shows the percentage of risk factors affecting human body in slow progression.

As we deal with regression analysis which is predictive based data mining technique, target values are considered and then from it predictors are chosen. In the above table age is chosen as target value and NCD risk factors are considered as predictors. For clear understanding a graph is plotted which explains based on the age and NCD risk factor the percentage of chronic diseases affecting human body in slow progression. The higher the percentage the higher is the affecting percentage in human body.

**II. LITERATURE REVIEW**

During past decades [12], efforts have been made to deal with these chronic diseases. Many of the risk factors for these heart disease, cancers, and stroke have been identified. Smoking known to be major health problem for the past and present decade which is finally being attacked in the developed world and lifestyle factors particularly nutrition(diet) and physical activity are slowly being tackled. Many of the risk factors for NCD are due to lifestyle and can be prevented, however decades are gone regarding the approaches to begin the work in developed countries and developing countries on tobacco companies but fast food chains are targeting developing countries. Drug treatments can be inexpensive but still remain out of reach for many people. Even when affordable, care is often not organized to deal with NCDs and their risk factors. The fight against non-communicable diseases needs to start today and be the prime focus tomorrow. The current work on smoking is a start but much more is needed.

In this paper, regression analysis [14] which is a predictive base data mining technique is considered .The regression is a supervised mining function for predicting a continuous, numerical target. The task of regression analysis begins with a dataset in which target values are known and user should predict the values (predictors). Regression analysis is a process of estimating the value of continuous target (Y) as a function (F) of one or more predictors ( $X_1, X_2, \dots, X_n$ ) and a set of parameters ( $\theta_1, \theta_2, \dots, \theta_n$ ) and a measure of error e, given by

$$Y = F(X, \theta) + e$$

Where, Y = target, is the dependent variable,

$X_1, X_2, \dots, X_n$  are predictors or independent variables,

$\theta_1, \theta_2, \dots, \theta_n$  are parameters or regression coefficients and

e represents error which is the residual (difference between expected & predicted value of dependent variable).

From table 1, 'age' is considered as target which is dependent variable and risk factors like tobacco use, physical inactivity, alcohol consumption and unhealthy diet are marked as predictors which are independent variables. Based on the risk factors the regression model can be created using Oracle Data Mining Tool and the regression model has to undergo testing by applying to test data with known target values and comparing the predicted values with the known values.

OBJECTIVE	METHOD/TECHNIQUES USED	DATASETS USED	ANALYSIS	PROS & CONS	REF
<p>To examine different machine learning approaches to use situational features associate with having or not having urge to smoke during a quit attempt.</p>	<p>Naïve Bayes classification, Discriminant analysis, decision tree learning methods.</p>	<p>Dataset collected from over 300 participants who had initiated a quit attempt.</p>	<p>The decision tree method/ classification tree method outperformed than the naïve bayes &amp; discriminant analysis method with accuracy up to 86%.</p>	<p>Machine learning classifiers help to identify smoking situations &amp; algorithms performance improves significantly based on best features &amp; classifier parameters. Cannot provide adaptive &amp; personalized support in minimum amount of time using novel expert systems.</p>	<p>15</p>
<p>To determine risk factors that characterize smoking behavior among older adults by assessing psychological distress, physical health status, alcohol use &amp; demographic variables in relation to current smoking.</p>	<p>An efficient data mining model CART (classification &amp; regression tree) that predicts smoking behavior of older adults with a minimal set of modeling assumptions.</p>	<p>Data collected from national survey on drug use and health(NSDUH,2006) which provides national alcohol &amp; drug use estimates for US civilian population annually. The NSDUH conducted subject interviews in person by responding to questions via a computer assisted interview module.</p>	<p>By using the techniques the explanatory variable used for prediction of average number of cigarette use per day when first started smoking cigarette is age followed by education level &amp; psychological distress.</p>	<p>CART is a data mining tool that provides hierarchical structure and facilitates interpretation of output assessment and decision making process in relation to program design. CART easily interprets statistical results. CART should be used with caution as the thorough understanding of theory and its guidance is critical to have a model to be tested with CART.</p>	<p>16</p>
<p>To classify smoking status of people considering second-hand smoking associated attributes: Allowance to second hand smoking at home, second hand smoking at work, exposure to smoke at work in the past 30 days.</p>	<p>Classification is performed using multilayer perceptron which is a well known neural network approach.</p>	<p>Data for analysis is obtained from Global Adult Survey (GATS) turkey. It aims at collecting data about tobacco use and tobacco control measures for adults aged 15 years and older. GATS is one the most used and control related surveys.</p>	<p>Based on the analysis the multilayer perceptron could correctly classify the smoking status of people over 68% using 5 attributes and can be interpreted as sufficient for this kind of problems.</p>	<p>The study analyze the capability of the classification of smoking status using 5 attributes and the accuracy level obtained is not so high, but is enough when the complexity of problem is considered. The above interpreted accuracy level is sufficient to take initial step for advance studies.</p>	<p>17</p>
<p>To classify a complex health behavior, smoking cessation between current and former smokers in 2000 (NHIS) National Health Interview Survey sample adult file.</p>	<p>Back Propagation neural network (BPNN) classifier is used for decision support systems in smoking cessation interventions.</p>	<p>The data is collected from NHIS corresponding to the adult core &amp; cancer control module are contained in 2000 NHIS sample adult file and can be downloaded from <a href="ftp://ftp.cd.c.gov/pub/Health_Statistics/NCHS/Datasets/NHIS/200/">ftp://ftp.cd.c.gov/pub/Health_Statistics/NCHS/Datasets/NHIS/200/</a> The 2000 NHIS sample adult file is a large dataset containing many health related attributes.</p>	<p>The BPNN classifier performance exceeded randomly with 95% asymmetric confidence intervals for area under receiver operating characteristic curve. Separation of current &amp; former smokers was imperfect as illustrated by the receiver operating characteristic curve (ROC) and performance of classifier did not exceed comparison classifier created using logistic regression.</p>	<p>To predict and model a health behavior, smoking cessation we consider back propagation neural network using health survey data. Performance of BPNN classifier did not exceed logistic regression model and poor data quality &amp; missing data in clinical datasets create a vast difference in classification scenario.</p>	<p>18</p>

<p>To find the performance of different classification methods of large datasets in heart disease prediction system. The algorithm which gives low error rate will be preferred as it has more powerful classification capability.</p>	<p>The algorithms compared are Naïve Bayes, K Nearest Neighbor, decision tree, artificial neural network.</p>	<p>The clinical dataset is considered which has parameters based on risk factors as age, family history, diabetes, hypertension, high cholesterol tobacco smoking, alcohol intake.</p>	<p>Among algorithms considered K Nearest Neighbor has improved classification method for dataset of heart in medical and bioinformatics field.</p>	<p>The best classification algorithm based on heart data is K Nearest Neighbor classifier and has an accuracy of 100%. The different accuracy % depends upon number of attributes taken used for implementing heart disease patient's availability of huge amount of data.</p>	<p>19</p>
<p>To analyze influence of NCD in study population by predicting probability of developing NCD depending on their lifestyle habits by implementing awareness programs.</p>	<p>K-means Clustering Data mining techniques is used.</p>	<p>As study population is considered it subjects to 303 employees working in public sector company in Mumbai region and data is collected from 2009-2011, the database involves complete information.</p>	<p>The datasets considered are diabetes, hypertension and dyslipidemia and applying temporal cluster graphs it was observed that diabetes (23.43%), hypertension (23.1%), dyslipidaemia (20.62%).</p>	<p>Data mining techniques can predict diseases fairly by deciphering their lifestyle habits. This improves individual quality of life by preventing long term complications and increasing life expectancy.</p>	<p>20</p>
<p>To provide an efficient way to extract the required clinical information form voluminous, raw and heterogeneous data. To provide techniques that predicts and forecast medical diseases with high accuracy and low cost.</p>	<p>For finding accuracies on various diseases commonly used techniques are (Dtrees) Decision trees, (ANN) Artificial Neural Network and Naïve Bayes(NB).</p>	<p>Based on various diseases authors have used datasets available online or form hospitals or respective centers. For drug abusers no datasets are available it needs an interview individually with each patient at drug addiction centers.</p>	<p>ANN shows maximum accuracy of 93.9% and minimum accuracy of 67.6%. Decision Tree has given maximum accuracy of 99.2%. Naïve Bayes has given maximum accuracy of 96.5% in curing heart diseases. Accuracy is major constraint in medical field which fails with minor fluctuations, more accuracy better are the results.</p>	<p>The algorithms which are considered influence the modifications of data values which are discovered on statistical importance of patterns which would be impossible to develop a universal measure for all data mining algorithms. The bigger problem is even if data mining results are credible, convincing the health practitioners need to change their habits based on their evidence.</p>	<p>21</p>
<p>To calculate the prevalence of 12 chronic outpatient diseases through characterization of outpatient prescriptions.</p>	<p>Classification techniques used are Decision tree, Support vector machine, Naïve Bays, Neural Network, Logistic regression and Nearest Neighbor.</p>	<p>The datasets includes 1412 prescriptions that are provided by Iran's Ministry of Health, Food &amp; drug administration.</p>	<p>Based on the classification techniques considered Naïve Bays has accuracy of 16% worse than Nearest Neighbor method which has lowest level of accuracy among other classification algorithms.</p>	<p>Data mining algorithms are suitable for classification problems.</p>	<p>22</p>

**Fig.2 objectives of authors working in the field of predicting and classifying smoking cessation status using data mining methodology**

The figure below highlights the objectives of authors working in the field of predicting and classifying smoking cessation status using data mining methodology. The objectives of many authors reveal different results on smoking cessation and influence of NCD in human population. Different data mining techniques are chosen based on the datasets used and performs with better accuracy. The bigger problem is even if data mining results are credible and convincing; the health practitioners need to change their habits based on the evidence. The data mining techniques can predict slow progression of

chronic diseases by deciphering their lifestyle habits. This improves individual quality of life by preventing long term complications and increasing life expectancy. Each objective of different authors is met by comparing and by analyzing accuracy of each technique. But the previous work which was done did not use e regression analysis which is a predictive data mining technique. This paper mainly talks about predictive analytics as it is today's challenging task in many organizations to analyze current and historical facts in order to make predictions about future events.

### III. METHODOLOGY

We use Oracle Data Miner (ODM) [23][24][27], a software mining tool for tracking smoking analysis report. ODM uses SVM [26][28] for regression and other mining functions. It supports our algorithms for computing regression, which are suited for mining high dimensional and unstructured datasets.

#### ORACLE DATA MINING

ODM has several data mining and data analytics algorithms for classification, prediction, regression etc. The data mining algorithms, which are parts of the oracle relational database directly join the oracle database kernel and work on the data that is stored in the relational database tables. The advantage of ODM is that it eliminates the need for extraction or transfer of data into analytic/mining servers. It provides a GUI which is an ODM tool and allows the user in the process of creating, testing and applying models on data in the process of creating, testing and applying models on data. ODM helps machine learning approaches such as per planes decision tress, naive Bayes, support vector machine (SVM), generalized linear model (GLM) which are used in predictive mining whereas association rules, k-means clustering and non-negative matrix factorization used for descriptive mining.

#### SVM [26][28]

SVM performs classification of data by finding the hyper plane that maximizes the margin between two classes. The classification is done using binary classifier and the data is linearly separable to generate two groups, of which one is positive and the other one is negative. The main objective of SVM is to find the best separating hyper plane which is also referred as decision boundary. It contains inputs/points which are vectors P, such that  $P = (X_1, X_2, \dots, X_n)$ . SVM decides the best separating line, which is the line that bisects and is perpendicular to the connecting line.

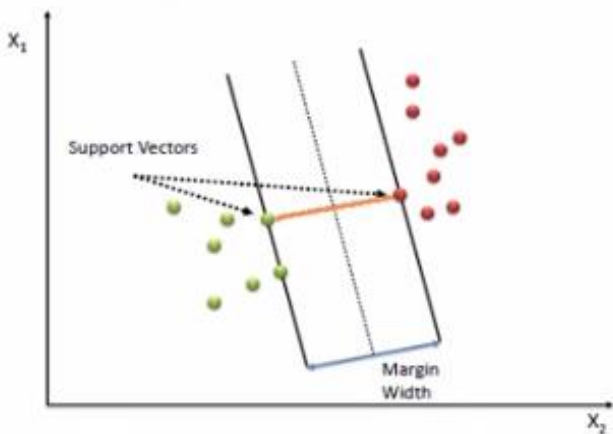


Fig.3 classifying data using hyper plane in vector space

Dataset can be represented as

$$D = \{(X_1, Y_1), (X_2, Y_2), \dots, (X_n, Y_n)\}, \text{ where}$$

$n$  = no of instances,  $X_i = i^{\text{th}}$  point/vector and  $Y_i =$  class associated with that point and can be either -1 or +1.

The dot product is done between two vectors A and B, where  $A = (a_1, a_2, \dots, a_n)$  and  $B = (b_1, b_2, \dots, b_n)$  is given by

$$A \cdot B = a_1 b_1 + a_2 b_2 + \dots + a_n b_n$$

A separating hyper plane is expressed in terms of 'b' and a normal vector  $\vec{w}$  which is perpendicular to the hyper plane and is referred as weight vector. All points  $\vec{x}$  on hyper plane satisfies the equation of the hyper plane  $(\vec{w})^T \cdot \vec{x} = -b$ . The classifier becomes  $f(\vec{x}) = \text{sign}((\vec{w})^T \cdot \vec{x} + b)$

In order to separate two groups of data a straight line or hyper plane of linear separability can be chosen. There are groups where a nonlinear region can separate the group more efficiently. SVM handles this nonlinear region data using a kernel function to map the data into a different space where a

hyper plane cannot be used to do separation. It means a nonlinear function uses linear learning machine in a high dimensional feature space while the capacity of the system is controlled by a parameter that does not depend on the dimensionality of the space which is called "kernel trick". The kernel function transforms the data into higher dimensional feature space to make it possible to perform linear separation.

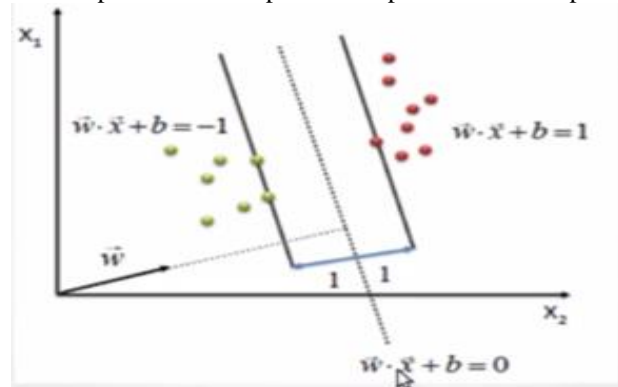


Fig.4 Linear classifier that is perpendicular to hyper plane

### IV SYSTEM DESIGN

The figure below explains system design of smoking cessation



Fig.5: working process of smoking cessation

The figure explains the working process of smoking cessation which requires 6 steps to be processed to achieve reducing/cessation tobacco use which is a key component in national or global plan.

**STEP: 1 CUSTOMER**

- 1) If a customer wants to buy a cigar he/she needs to use Aadhaar card and the transaction should be done based on it.
- 2) The customer should insert Aadhaar card into the KIOSK machine and follow further proceedings.

**STEP: 2 AADHAAR CARD**

Customer should use his/her Aadhaar card if he wants to purchase a cigar as the procedure is linked using Aadhaar card.

**STEP: 3 KIOSKS**

A kiosk [29] is a small physical structure (often including a computer and a display screen) that displays information for people. Kiosks let users interact and include touch screens, sound, and motion video. The display information regarding purchasing of cigar is purely based on AEPS (Aadhaar enabled payment system). The AEPS system should work as back end in KIOSK and it provides the information from Aadhaar

database regarding the person who ever uses Aadhaar card to purchase a cigar. Conditions are imposed on the customer when he purchases cigars. There will be a daily limit for purchasing cigars, if the limit exceeds the customer will not be permitted to do any purchasing transactions. The tax rate will be imposed on each cigar. All the above mentioned conditions and facilities will be made available on the KIOSK using AEPS.

AEPS stands for Aadhaar Enabled Payment System[30] which is a way to get money from bank account. The UPI (unified payment interface of India) [31] is a back end for AEPS and it is a payment system which facilitates the fund transfer between two accounts and works on mobile platform. You are not required to give bank account details for the amount to be transferred or amount to be paid when purchasing a cigar through the UPI payment system. UPI is one of the most advanced methods among all the digital payments.

**STEP: 4 AADHAAR DATABASE**

When customer tends to buy cigar and the transaction is done using Aadhaar card of the customer which will be done by AEPS in KIOSK. Now there will be any entry in the database of the customer which conveys the information regarding how many cigars have been purchased throughout the month and details of the customer like name, age, address, contact number, and the bill generated for particular month, tax that has been deducted for consuming cigars for the particular month.

**STEP: 5 ODM GENERATED HEALTH REPORT**

Based on the data which has been traced from Aadhaar database, an Oracle Data Miner based health report [23][24][27] will be generated for each customer who has purchased cigar. The report gives a prediction analytics about the health condition of the customer and it forecasts what might happen in the future with an acceptable level of reliability, and how far the person has been affected with Non communicable risk factors, and his/her health condition. The report includes what-if scenarios & risk assessment based on the age group.

**STEP: 6 CONTACT DETAILS /NUMBER OF CUSTOMER**

The report will be delivered to the customer mobile as message based on the details provided on Aadhaar card and available in Aadhaar database.

**V. MAJOR FEATURES OF THE SYSTEM**

The proposed system has the following features.

- Reducing/cessation tobacco use is a key component in national or global plan.
- Generates a Health report of tobacco adductors/smokers.

- Increases tax rate for each cigar which is beneficial for government, and tax is deducted from person/individual salary.
- As predictive analytics is being used, it forecasts what might happen in the future with an acceptable level of reliability.
- Includes what-if scenarios and risk assessment.
- Linking the generated health report to government welfare programs as it is not a natural health hazard.
- Instead of selling cigars in market we go for KIOSK which includes a computer & display screen that displays information for people) that should work based on national payment corporation of India that can be AEPS (Aadhaar enabled payment system).
- Collect insurance from customer when he buys a cigar which benefits government, and the insurance should be hidden or made transparent from customer/user where it can be prevention activity not promotional activity.
- We can go for less usage of tobacco than we can prefer herbs which reduces the victim to stop smoking.
- Encourage the victim when he stops smoking based on the monthly generated health report.
- Based on age group treatment/cautions are given to patients immediately or delayed.
- If there is no improvement in health report of victim for the coming months then he/she need to pay penalty for the government.

#### **SOME CONSTRAINTS FOR THE SYSTEM**

- As the investment for KIOSK is more it can be a drawback which cannot be succeeded at rural areas.
- When an individual buys a cigar & sells it to other person/third party, the health report which is generated will go waste.
- The report which is generated may be refused or may not be considered with much concern for the first few months or perhaps for a year.
- But later on it might become an advantage to the person whoever it belongs.
- If the victim is an illiterate and doesn't know what the report actually says then the report is useless.
- As we sell cigars based on Aadhaar card and if the details of the customer are wrong then the generated report will be a cause not to cease smoking.
- Collecting data from Aadhaar database may not be permitted by the government as they are human rights.

#### **VI CONCLUSION**

As our main focus is on predictive analysis of smoking, which is a non-communicable disease it forecasts what might happen in the future with an acceptable level of reliability and includes what-if scenarios and risk assessment. Based on the age group

different (treatments or cautions) are given for patients immediately or delayed. Thus reducing tobacco use must be a key component in national or global plan. This study provides measures in reducing tobacco use.

#### **REFERENCES**

- [1]<http://www.who.int/bulletin/volumes/90/1/11-091041/en/>
- [2] <http://www.who.int/gho/database/en/>
- [3] <http://apps.who.int/gho/data/node.home>
- [4]<http://www.tobaccoatlas.org/topic/non-communicable-diseases/>
- [5] <http://ghdx.healthdata.org/series/steps-noncommunicable-disease-risk-factors-survey>
- [6]<http://www.who.int/mediacentre/factsheets/fs355/en/>
- [7] [https://en.wikipedia.org/wiki/Tobacco\\_smoking](https://en.wikipedia.org/wiki/Tobacco_smoking)
- [8] <http://www.who.int/topics/tobacco/en/>
- [9]<http://www.who.int/mediacentre/factsheets/fs385/en/>
- [10]<http://www.healthline.com/health/alcoholism/basics#Overview>
- [11]<http://www.healthline.com/health/alcoholism/basics#Overview>
- [12] Alberti, G.: Noncommunicable diseases: tomorrow's pandemics, Bull World Health Organ vol.79 n.10 Genebra, Jan. 2001
- [13]<http://www.wikihow.com/Run-Regression-Analysis-in-Microsoft-Excel>
- [14][https://docs.oracle.com/cd/B28359\\_01/datamine.11/b28129/regress.htm#DMCON005](https://docs.oracle.com/cd/B28359_01/datamine.11/b28129/regress.htm#DMCON005)
- [15]Dumortier, A.,Beckjord, E.,Shiffman, S. and Sejdic,E.: Classifying smoking urges via machine learning, 2016 Elsevier Ireland Ltd-Sep-2016
- [16] Moon, S. S., Kang,S. Y., Jitpitaklert, W. and Kim, S. B.: Decision tree models for characterizing smoking patterns of older adults,2001 Elsevier Ltd.
- [17]Durmusoglu, Z. D.U.,Ciftci,P.K.: Classification of Smoking Status: The Case of Turkey,2016 IEEE
- [18]Poynton, M. R., McDaniel,A. M.: Classification of smoking cessation status with a back propagation neural network,2006 ElsevierInc.
- [19] Tamilarasi,R. and Porkodi,R.: A study and Analysis of Disease prediction techniques in Data Mining for Healthcare,IJERMt-Marh-2015
- [20]Waghmare, K. J.: Prediction of occurrence of Non-communicable Diseases depending on Lifestyle Habits using Clustering Data Mining Technique,Elsevier publication-2013.
- [21] Kaur, S. and Bawa,R.K.: Future trends of Data mining in predicting the various diseases in Medical Healthcare sytem,IJEIC-2015
- [22]Alamdari, M. S.,Teimouri, M.,Farzadfar, F. and Meshkini, A. H.: Disease detection in Medical Prescriptions using data mining tools,2014 IEEE
- [23]<http://www.oracle.com/technetwork/database/options/advanced-analytics/odm/odminstallation-2080768.html>



- [24]<http://www.oracle.com/webfolder/technetwork/tutorials/obe/db/11g/r2/prod/bidw/datamining/ODM11gR2-SetUp.htm>
- [25]<http://www.predictiveanalyticstoday.com/what-is-predictive-analytics/>
- [26][http://docs.opencv.org/2.4/doc/tutorials/ml/introduction\\_to\\_svm/introduction\\_to\\_svm.html](http://docs.opencv.org/2.4/doc/tutorials/ml/introduction_to_svm/introduction_to_svm.html)
- [27][https://en.wikipedia.org/wiki/Oracle\\_Data\\_Mining](https://en.wikipedia.org/wiki/Oracle_Data_Mining)
- [28]<https://www.analyticsvidhya.com/blog/2015/10/understanding-support-vector-machine-example-code/>
- [29] <http://businessjargons.com/kiosk-banking.html>
- [30] <https://www.quora.com/What-is-Aadhaar-enabled-payment-system>
- [31] <https://upipayments.co.in/upi-payment-system/>