ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 4, April 2017

Health Analytics Using Machine Learning: A Survey

Apoorva Sharma[1], Pallavi Rawat[2], Kajal Pandey[3], Ravi Shankar Rai[4]

B.E. Student Department of Computer Engineering, Army Institute of Technology, Pune, India

ABSTRACT: We present a method to analyse the risk of diseases like diabetes, sexually transmitted infections, increased blood pressure, breast cancer, heart disease and chronic kidney disease based on the patient symptoms, past diagnosis and lifestyle. The presented methodology may be incorporated in applications like communication and decision support systems in health care, risk management, health analysis and disease prevention.We use datasets of different diseases from sources like data.gov.in, UC Irvine(UCI) machine learning datasets, pimaindian diabetes data, etc. When a user enters the symptoms related to a disease we classify the patient in one of the disease categories. Then taking the patient lifestyle in account, we analyse the degree of risk for the particular disease.We use Naive Bayes classifier and C4.5 decision tree to classify patients in various categories. These classifiers can also be compared with other classifiers like logistic regression, artificial neural networks, support vector machines, random forests, bagging and boosting.In case of high dimensional data, it can be reduced using principal component analysis (PCA) and random sub sampling. The method we proposed will predict accurate analysis of patient data.

KEYWORDS: Machine learning; healthcare analytics; classification algorithms; decision tree; naïve bayes; Apache Hadoop; Apache Spark

I. INTRODUCTION

Conventional medicine requires doctors and other health care professionals to treat diseases using drugs, radiation and therapy. These professionals are well trained in the field of medicine. But it is not possible to remember all the information that they may need for every circumstances. Even if the professionals had access to all the data that they needed to treat the diseases they face, it would take a long time for them to analyse all of that data and come up with a suitable solution based on the patient's medical profile. Predictive analytics uses methods to read the huge data, analyse it and predict consequences for patients. The data has historical as well as real time data. The historical data takes into account the past treatment outcomes of the patient. The real time data includes the latest trends in treatment. This large amount of information cannot be dealt with by even a human expert for every patient. It is understandable that the past diagnostic history of a patient can present a good opportunity to understand the nature of the disease. Also, the past treatment can explain what went right and what did not go as expected. This may be different for every patient. The present condition of the patient could be a reaction to his past treatment. There may be now new trends in the industry which may not have been utilised before and can provide significant enhancement to the treatment. The health care industry needs to deal with many problems related to cost and quality. The problems can be dealt with if institutions decide to incorporate prescriptive analytics. Prescriptive analytics does not only show a result which may occur but also suggests how health care can become more patient need oriented. A model such as this will be helpful in many ways. The treatment can be improved and cost of health care can be reduced. This model, however, cannot replace human involvement. It can only provide a way for physicians to support their decision making to present the best results. Advantages of using machine learning in health care are

- More accurate diagnosis.
- Early involvement to prevent diseases.
- If the predicted risk is high, necessary steps can be taken to avoid the disease.
- Patients can use this system for information for self.

We aim to analyse the risks of diseases like diabetes, breast cancer, sexually transmitted infections, increased blood pressure, heart disease and chronic kidney disease in individuals based on their diagnostics history, symptoms

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 4, April 2017

experienced and current lifestyle using machine learning and to suggest preventive measures, assessment and for information to the patient.

There has been a lot of work over the years to develop a model that can be used to predict the risk of various diseases in individuals in order to prevent them and reduce the risk. Most diseases may depend on the past of the patient and his life style may affect the chances of getting one or more diseases.

II. RELATED WORK

Machine-learning technologies and predictive analytics have been utilized for decades across a number of industries. In recent years, the healthcare sector has begun adopting these technologies for a variety of applications, including chronic disease management, staffing predictions, population health risk assessment and for information to the patient.

Different papers published over the years have tried to develop a system to predict the risks of various diseases. Work has been done to develop classification models using various algorithms like naive bayes, C4.5 decision tree, random forest, artificial neural networks, etc. The algorithms have been found to provide different percentage of accuracy where some have proved to be better than others. These have been discussed below.

Year	Publication	Author	Title	Algorithms	Conclusion	Limitations
2015	Engineering in Medicine and Biology Society (EMBC)	A.Voss, R.Shroeder, M.Vallverdu	Linear and non-linear heart rate variability risk stratification in heart failure patients &	Non linear symbolic dynamics	HRV measures and other parameters higher risk of heart failure. These measures are taken from non linear dynamics	The results do not depend on what caused the heart failure. Future experiments needed to verify this by additional studies.
2016 &	IEEE Transactions on Multimedia	Ahmed M. Alaa, Kyeong H. Moon, William Hsu	ConfidentCare: A Clinical Decision Algorithm Support System for Personalized Breast Cancer Screening	Supervised learning, C4.5 Decision Algorithm	Algorithm creates cluster and learns from each cluster. The clusters are generated based on features iteratively	Needs personalised attributes of patients and does not handle missing values.
2015	TENCON IEEE &	Lakshmi B.N., Indumathi T.S., Nandini Ravi	A comparative study of classification algorithms for risk prediction in pregnancy	C4.5 Decision Tree Classification Algorithm, Naive Bayes	C4.5 decision tree has greater potential in accuracy for predicting the risk levels during pregnancy.	Other classifiers were not considered for this study.
2015	IEEE Journal of Biomedical and Health Informatics	Bum Ju Lee, Jong Yeol Kim	Identification of Type 2 Diabetes Risk Factors Using Phenotypes Consisting of Anthropometry	Naive Bayes, Logistic Regression	Waist circumference was a better predictor of risk of diabetes than triglycerides.	The phenotypes are considered for certain ethnicities and not for

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 4, April 2017

			and Triglycerides based on Machine Learning			the general population of the world.
2015	Engineering in Medicine and Biology Society (EMBC)	Yajuan Wang, Kenney Ng, Roy J. Byrd	Early detection of heart failure with varying prediction windows by structured and unstructured data in electronic health records	Predictive HF model	As the prediction window decreases, the performance of the model increased.	The prediction percentage of unstructured data was less than that of structured data.
2015	Engineering in Medicine and Biology Society (EMBC)	OgnjenArandjelovi	Prediction of health outcomes using big (health) data	Bottom up modelling, Direct high-level modelling	The future of a patient can be predicted from his past state. This depends on his present value of various attributes.	Markov process-based model performs better in 18% of the cases.
2014	International Journal of Computer Science and Information Technologies	MukeshKumari, Dr.Rajan Vohra, Anshul Arora	Prediction of Diabetes Using Bayesian Network	Bayesian network	Classification with Bayesian classifier shows the best accuracy for diagnosis of diabetes.	All risk factors have not been considered. Bayesian classifier is not sufficient when there are missing values.
2014	BMC Medical Informatics and Decision Making	Mohammed Khalilia, Sounak Chakraborty, MihailPopescu	Predicting disease risks from highly imbalanced data using random forest	Repeated random subsampling, Support vector machine, bagging, boosting, random forest	In combining repeated random sub-sampling with RF overcame the class imbalance problem to predict diseases.	Difficulty in accessing full medical records due to privacy issues. The dataset was highly imbalanced. Duplicate data.
2014	Journal of Obesity	Hudson FernandesGolino, Liliany Souza de Brito Amaral, Stenio Fernando	Predicting Increased Blood Pressure Using Machine learning	Classification and regression tree (CART)	For women WC, BMI and WHR provided more accurate results. For men WC, HC, WHR and BMI together presented more accurate	Variance issue: this means that the algorithm learned too much from the test data and is likely

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 4, April 2017

					results than BMI alone.	to make more errors in a different dataset.
2015	Biomedical Informatics Training, Stanford University	Linbailu Jiang, Yufei Zhang, Siyi Peng	Data Fusion for Predicting Breast Cancer Survival	SVM, Naive Bayes, 10-fold cross validation	It has much higher accuracy on predicting the patient's survival status than simply treating the whole problem as a classification model and implementing support vector machine or Naive Bayes model.	The threshold was set as 0.5. It can be raised to increase specificity without decreasing the sensitivity too much.
2015	Biomedical Informatics Training, Stanford University	William Chen, Henry Wang	Predicting Breast Cancer Survival Using Treatment and Patient Factors	SVM, recursive partitioning, random forest, gradient-boosted classification tree	The survival of a patient after five years is fairly consistent with the overall set of features given but there is a small group of drugs/treatments that is extremely predictive. In fact, the subset of ten treatments found is enough to make predictions that are about as accurate as using the entire feature set.	Dataset cannot be explained by a linear data model. It can be improved to find those at a greater risk than others.
2014	Journal of Machine Learning, Stanford University	Predicting Heart Attacks	Sihang Yu, Xuyang Zheng, Yue Zhao	Multi-class supported vector machines (SVM), Multi-class Naive Bayes (NB), decision tree, random forest	The test accuracy of random forest is significantly better than other models. To make a prediction, all of the models in the ensemble are polled and their results are averaged.	Small amount of data and missing data. The model does not work on real-time data such as ECG signals.
2015	Journal of Machine Learning, Stanford University	Junrui Zhang, Duyun Chen	Methods for predicting Type 2 diabetes	Logistic Regression, SVM, Random Forest,	Balancing the data set can improve the prediction, and oversampling generally works	The dataset is large and diverse. It requires special clean

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 4, April 2017

				Decision Tree	better than undersampling	up methods and better feature selection through domain knowledge.
2014	Biomedical Informatics Training, Stanford University	Maulik R. Kamdar	Visualizing Personalized Cancer Risk Prediction	Gaussian Naive Bayes, SVM, Decision tree, Ensemble method random forest	Better evaluation metrics and PCA clusters are obtained for classifiers trained using DM Data.	Only SVM classifiers provided desirable specificity and sensitivity. Other pairs generated skewed pairs.

Table.1. Literature Survey

III. ARCHITECTURAL DESIGN

A. *Design Considerations:*

The system is designed as a three-tier architecture consisting of a front end, a back end and a database.

- Front End: The front end is a web application based graphical user interface in which the user can specify symptoms and other demographic details.
- Back End: The back end is an analytical model designed using machine learning to analyse the risk of diseases based on the symptoms and other demographic details mentioned.
- Database: The database is a collection of datasets of various diseases for which the risk is being analysed.

Fig 1 shows the basic architecture of the system. The lower most layer is the data layer. Data are stored in Hadoop Distributed File System (HDFS). Next is the processing layer which uses various classifying algorithms to create a model. The top player is the graphical user interface layer which allows user to interact with the system easily.

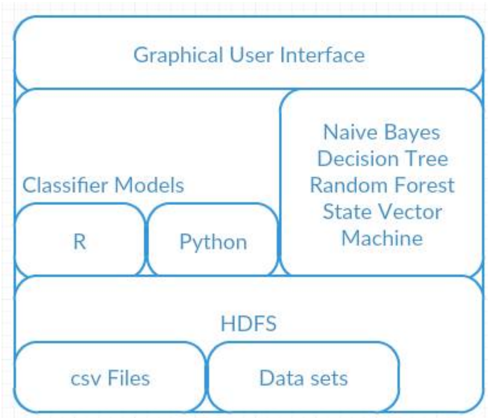

Fig.1. Basic Architecture

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 4, April 2017

Fig 2 shows an applied conceptual architecture of analytics. It describes how the data stored in the database are first transformed using middleware. These transformed data are then stored in Hadoop. The big data analytics tools access the data from Hadoop and use machine learning classifiers to create a model. This model can then be used for prediction.

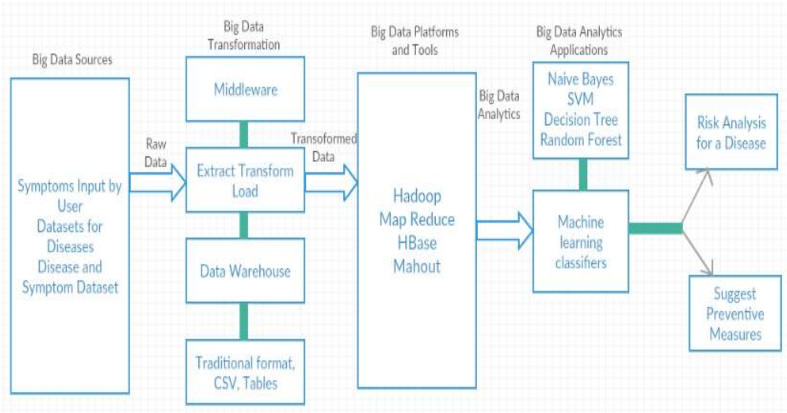

Fig.2. An Applied Conceptual Architecture of Analytics

B. *Description of the Datasets:*

- Diabetes Database: This dataset was taken from machine learning archive of UCI. The original owners are National Institute of Diabetes and Digestive and Kidney Diseases. The donor of database is Vincent Sigillito. The data set was collected from women who are at aleast 21 years of age and belong to pimaindian heritage. Total number of rows are 768. Total number of features are 8 plus class attribute.
- Women's Sexual Health: The data was collected from around 9000 young (15 to 30 years old) woman subjects when they visited clinics in 9 underdeveloped regions, with around 1000 subjects in each region. Each subject was asked by clinical practitioners some questions and her answers were recorded, together with her demographic information. The sexual and reproductive health risks were then evaluated by clinical practitioners and are assigned to different risk segments and subgroups.
- Blood Pressure Dataset: This dataset was obtained from a study that tried predicting increased blood pressure by using different features. Data were collected from college students, both male and female.
- Breast Cancer Diagnostic Dataset: The attributes are the characteristics of the cell nuclei present in the image of mammogram. These are used to predict whether the tumor is malignant or benign.
- Heart Disease Dataset: This dataset was also obtained from UCI machine learning archive. The directory contains different databases for heart disease. We have selected the dataset for Cleveland.
- Chronic Kidney Disease: This dataset was taken from UCI Machine Learning datasets. It contains 400 instances. The number of attributes is 25.

IV. CONCLUSION AND FUTURE WORK

Predictive analytics is the most discussed topic when it comes to health care analytics. Machine learning is a discipline that has been studied well and has a long history of success in various fields. Health care can make use of the previous success and learn lessons to start using predictive analytics for improving various issues related to health care. These issues include improving patient care, chronic disease management, hospital administration and supply chain efficiencies. The health care systems need to understand what predictive analytics means to them and how it can be used most effectively to improve their system.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 4, April 2017

Prediction can be used in a most efficient manner if that knowledge can be transferred into action. Therefore, it requires the willingness to intervene to make best use of historical and real time data. Under value-based care models, providers must proactively manage the health of individuals with chronic illness to curtail costly complications that can lead to hospitalization, hospital readmission and/or early death. Many chronic diseases are linked to unhealthy behaviors, such as lack of physical activity, tobacco use and poor nutrition. Providers are motivated to closely monitor these behaviors and take action to keep patients healthy. Our proposed system will help users know what disease the symptoms point at and how high is the risk of the user to have the illness. Since, we consider the user's demographic details and lifestyle, we can suggest ways to lower the risk. In this way we present an effective health analytics system using machine learning.

There is still a lot of work to be done in this field that can improve the accuracy for disease prediction. For some diseases the data available is not enough to design a classifier model that can make prediction for disease control. Also the healthcare prediction is not accurate enough that it can be depended upon. Our proposed system only covers a few diseases. This model can be expanded in future covering as many diseases as possible, so that not only a person can be diagnosed for any type of disease but is also provided relevant solution for the same like suggesting doctors for his disease or some home remedies etc.

REFERENCES

1. Ahmed M Alaa, Kyeong H Moon, William Hsu, and Mihaela van der Schaar. Con_dentcare: A clinical decision support system for personalized breast cancer screening. arXiv preprint arXiv:1602.00374, 2016.
2. BN Lakshmi, TS Indumathi, and Nandini Ravi. A comparative study of classi_cation algorithms for risk prediction in pregnancy. In TENCON 2015-2015 IEEE Region 10 Conference, pages 1{6. IEEE, 2015.
3. Bum Ju Lee and Jong Yeol Kim. Identi_cation of type 2 diabetes risk factors using phenotypes consisting of anthropometry and triglycerides based on machine learning. IEEE journal of biomedical and health informatics, 20(1):39{46, 2016.
4. Yajuan Wang, Kenney Ng, Roy J Byrd, Jianying Hu, ShahramEbadollahi, Zahra Daar, Steven R Steinhubl, Walter F Stewart, et al. Early detection of heart failure with varying prediction windows by structured andunstructured data in electronic health records. In 2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), pages 2530{2533. IEEE, 2015.
5. OgnjenArandjelovi_c. Prediction of health outcomes using big (health) data. In 2015 37th Annual nternational Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), pages 543{2546. IEEE, 2015.
6. MukeshKumari, Rajan Vohra, and Anshul Arora. Prediction of diabetes using bayesian network. 2014.
7. Mohammed Khalilia, Sounak Chakraborty, and MihailPopescu. Predicting disease risks from highly imbalanced data using random forest. BMC medical informatics and decision making, 11(1):1, 2011.
8. Hudson FernandesGolino, Liliany Souza de Brito Amaral, Stenio Fernando Pimentel Duarte, Cristiano Mauro Assis Gomes, Telma de Jesus Soares, Luciana Araujo dos Reis, and Joselito Santos. Predicting increased blood pressure using machine learning. Journal of obesity, 2014, 2014.
9. Max Drach. Predicting life expectancy of acute myeloid leukemia (aml)patients based on gene expression of cancer cells.
10. Albert Y Lui and Alexandra M Pappas. Thyroid dysfunction: Predictionand diagnostics. 2015.
11. Duyun Chen, Yaxuan Yang, and Junrui Zhang. Methods for predictingtype 2 diabetes cs229 final project december 2015.
12. Andreas Voss, Ste_en Schulz, Rico Schroeder, Mathias Baumert, and PereCaminal. Methods derived from nonlinear dynamics for analysing heartrate variability. Philosophical Transactions of the Royal Society of LondonA: Mathematical, Physical and Engineering Sciences, 367(1887):277{296,2009.
13. Tina R Patil and SS Sherekar. Performance analysis of naive bayes andj48 classi_cation algorithm for data classi_cation. International Journal ofComputer Science and Applications, 6(2):256{261, 2013.
14. Sona Taheri, John Yearwood, Musa Mammadov, and SattarSeifollahi. Attributeweighted naive bayesclassi_er using a local optimization. NeuralComputing and Applications, 24(5):995{1002, 2014.
15. Wei Dai and Wei Ji. A mapreduce implementation of c4. 5 decision treealgorithm. International Journal of Database Theory and Application, 7(1):49{60, 2014.
16. AmuthanPrabakarMuniyandi, R Rajeswari, and R Rajaram. Networkanomaly detection by cascading k-means clustering and c4. 5 decision treealgorithm. Procedia Engineering, 30:174{182, 2012.
17. HyungkwonKo, Jaehoon Sung, Sue Min Cho, and Taeseon Yoon. Comparisonof the performances of the decision tree algorithm c5 using roughset and the neural network. 2014.
18. Katherine Ellis, Jacqueline Kerr, SuneetaGodbole, GertLanckriet, DavidWing, and Simon Marshall. A random forest classifier for the predictionof energy expenditure and type of physical activity from wrist and hipaccelerometers. Physiological measurement, 35(11):2191, 2014.
19. Shan Suthaharan. Support vector machine. In Machine Learning Modelsand Algorithms for Big Data Classification, pages 207{235. Springer, 2016.
20. JasonWeston. Support vector machine. Tutorial, http://www.cs.columbia.edu/~kathy/cs4701/documents/jason svm tutorial. Pdf, accessed, 10, 2014.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 4, April 2017

21. Zhiquan Qi, Yingjie Tian, and Yong Shi. Robust twin support vector machine for pattern classification. Pattern Recognition, 46(1):305{316, 2013.
22. Hanspeter A Mallot. Arti_cial neural networks. In Computational Neuroscience,pages 83{112. Springer, 2013.
23. Filippo Amato, Alberto Lopez, Eladia Maria Pena-Mendez, Petr Vanhara, Ales Hampl, and Josef Havel.Artificial neural networks in medical diagnosis.Journal of applied biomedicine, 11(2):47{58, 2013.
24. David W Hosmer Jr, Stanley Lemeshow, and Rodney X Sturdivant. Assessing the fit of the model. Applied Logistic Regression, Third Edition,pages 153{225, 2013.
25. Ji Zhu and Trevor Hastie. Kernel logistic regression and the import vectormachine. Journal of Computational and Graphical Statistics, 2012.
26. NahitEmanet, Halil R Oz, NazanBayram, and DursunDelen. A comparativeanalysis of machine learning methods for classi_cation type decisionproblems in healthcare. Decision Analytics, 1(1):1, 2014.
27. Nathaniel Roysden and Adam Wright. Predicting health care utilizationafter behavioral health referral using natural language processing and machinelearning. In AMIA Annual Symposium Proceedings, volume 2015,page 2063. American Medical Informatics Association, 2015.
28. Cynthia Rudin and Kiri L Wagsta_. Machine learning for science andsociety. Machine Learning, 95(1):1{9, 2014.
29. WullianallurRaghupathi and VijuRaghupathi. Big data analytics inhealthcare: promise and potential. Health Information Science and Systems,2(1):1, 2014.
30. Jimeng Sun and Chandan K Reddy. Big data analytics for healthcare. In Proceedings of the 19th ACM mSIGKDD international conference on Knowledgediscovery and data mining, pages 1525{1525. ACM, 013.
31. KyoungyoungJee and Gang-Hoon Kim. Potentiality of big data in themedical sector: focus on how to shape the healthcare system. Healthcareinformatics research, 19(2):79{85, 2013.
32. Raghunath Nambiar, Ruchie Bhardwaj, AdhiraajSethi, and RajeshVargheese. A look at challenges and opportunities of big data analyticsin healthcare. In Big Data, 2013 IEEE International Conference on, pages17{22. IEEE, 2013.
33. https://archive.ics.uci.edu/ml/machine-learning-databases/.
34. Vincent Sigillito (vgs@aplcen.apl.jhu.edu). Research center, rmi groupleader applied physics laboratory the johns hopkins university johns Hopkins road laurel, md 20707 (301) 953-6231.

BIOGRAPHY

Apoorva Sharma is a final year Bachelor of Engineering student in the Department of Computer Engineering, Army Institute of Technology, Pune. Her research interests are machine learning and big data.

Pallavi Rawat is a final year Bachelor of Engineering student in the Department of Computer Engineering, Army Institute of Technology, Pune. Her research interests are machine learning algorithms, Apache Spark, etc.

Kajal Pandey is a is a final year Bachelor of Engineering student in the Department of Computer Engineering, Army Institute of Technology, Pune. Her research interests are algorithms, data structures, machine learning.

Ravi Shankar Rai is a final year Bachelor of Engineering student in the Department of Computer Engineering, Army Institute of Technology, Pune. His research interests are algorithms, data structures, etc.