ISSN: 2278 – 1323
International Journal of Advanced Research in Computer Engineering & Technology (IJARCET)
Volume 5, Issue 9, September 2016

Cardiac Risk Prediction Analysis Using Spark Python (PySpark)

G.Tirupati, Prof. K.Venkata Rao

***Abstract*-Cardiovascular disease is the acute disorder in the world today. Disease control and early diagnosis of disorder can prevent from death and other diseases. Several techniques have been developed for assessment of cardiac risk using structured and unstructured patient data. Coronary Artery Disease(CAD) is predominated disorder occurs due to several parameters such as cholesterol level, Blood pressure, sugar levels, smoking status, age and family history. Usually data is very crucial for prediction of the risk and data is available in many formats such as structured, semi structured and unstructured data, among the data formats unstructured data is vital and risk factor parameters are embedded in it. This work presents an automatic method, which extracts clinical, physical and other parameters from unstructured data and these are used for predicting the cardiac disease risk and analyzed risk prediction methods such as Framingham, Reynolds and Prospective Cardiovascular Munster (PROCAM) using spark with python ((PySpark). Study observes Reynolds risk prediction method shows high sensitivity and specificity than other methods.**

So Reynolds risk prediction method provides better screening tool for both men and women to know the cardiac diseases and helps the patients that CAD can be prevented and controlled. It also provides statistical data of these methods to researchers and organizations.

Keywords- Accuracy, Cardiac Risk, Prediction, PySpark, Sensitivity.

I. INTRODUCTION

Cardiovascular disease is the leading global cause of death, accounting for more than 17.3 million deaths per year, a number that is expected to grow to more than 23.6 million by 2030[1]. Coronary Artery Disease (CAD) is the most common type of heart disease and leading cause of death in both men and women. CAD happens when the arteries that supply blood to heart muscle become hardened and narrowed their inner walls. This build-up is called atherosclerosis. It is due to the build up of cholesterol and other material, called plaque, on as it grows, less blood can flow through the arteries. As a result, the heart muscle can't get the blood or oxygen it needs. This can lead to chest pain or a heart attack. So prevention and control of this is vital in health care. It can be determined by various prediction methods, which is useful for both patient and clinician. CAD risk assessment is part of various national and international guidelines [2], and various risk assessment methods are existed for different age groups. One of the risk assessments is Framingham technique, which gives the guidelines for determining the 10-year risk factor both for men and women in the age group 30-64. One of the studies presents a rule based procedure for how to assess the risk using unstructured electronic records using text mining system [3]. Hui Yang described hybrid system to automatically identify the risk factors for heart disease [4], but both the methods have certain limitations. Now a day's big data in health care is vital role in analyzing the patient data using different analytic platforms [5]. A hadoop map reduce framework was proposed to analyze the major diseases such as diabetes and other disorders [6].

A systematic approach was developed to enhance known knowledge-based risk factors with additional potential risk factors derived from data. Systemic approach to enhance known knowledge based risk factors with additional risk factors derived from data [7]. A Congestive Heart Failure (CHF) case finding algorithm was developed, tested and prospectively validated. The successful integration of the CHF case findings algorithm into the Maine HIE live system is expected to improve the Maine CHF care[9] and A survey has been done on how big data analytics plays a role in predict the emergency situations before it happens [10]. A study shows comparison between Framingham risk scores and Reynolds risk score prediction [11]. But both methods use manual process for extracting the parameters. Evaketole and tiira laatikainen presented a paper on how do different cardiovascular risk scores act in real life [12], in this work sensitivity and specificity of risk charts based on Framingham, SCORE and CVD risk score. But these risk scores were manually calculated. Another study sharmini and selvarajah and gurpreet kaur presented work on comparison of the Framingham risk, SCORE and WHO/ISH

***G. Tirupati**, Department of Information Technology, GVP College of Engineering for Women, Visakhapatnam, India.*

***Prof. K.Venkata Rao**, Department of Computer Science & Systems Engineering, Andhra University College of Engineering, Visakhapatnam, India.*

cardiovascular risk prediction models in an asian population [13].SCORE high model predict risk accurately in men but underestimated women. A work presents how Reynolds risk scores effects based on c - reactive protein and family history [14].

II. MATERIALS AND METHODS

A. Data

Data is vital for risk prediction and further analysis. Most of the data available in unstructured format it is also known as clinical notes (unstructured). Clinical notes contain rich and diverse source of information. Challenges for handling clinical notes **g**rammatical, short phrases, Abbreviations, Misspellings, Semi-structured information. Unstructured patient data (also known corpus) available or gets from the informatics for integrating biology & the bedside (i2b2) track2 for identifying the risk assessment. XML data file contains elements both text and attributes. These elements describe patient information about present and past status as well as physical and clinical parameters. Data gets from informatics for integrating biology & bedside (i2b2) contains 53 xml files, some files already having CAD (abnormal) and remaining are normal. Each data file contains patient details such as medication, laboratory results, medical history and personal information (age, weight). Figure 1 shows how analytics can be used in healthcare.

Fig. 1 Model of Health Care Analytics

CAD risk parameters such as personal information, laboratory results and medical history mined from these data files using natural language processing tool kit.

B. Methods

It is automatic cardiac risk prediction method, which extracts physical, clinical and family history, in which Framingham risk prediction is one of the rule based technique depends on patient age, total cholesterol, High Density Lipoprotein (HDL) cholesterol, systolic Blood Pressure (BP), Treatment for hypertension and smoking status. Each parameter has some score points and numbers of points are based on range of the parameter and again these are different for men and women. All points are added and final risk score is determined by the total number of points. Figure 2 shows data file, which contains text and tags.

automatic Technique , which extracts the physical and clinical parameters from data file. It contains gender may be male or female or M or F and age also represented either years or Y. Some of attributes , for exmaple smoking parameter can be extracted based on status as shown in fig.2.

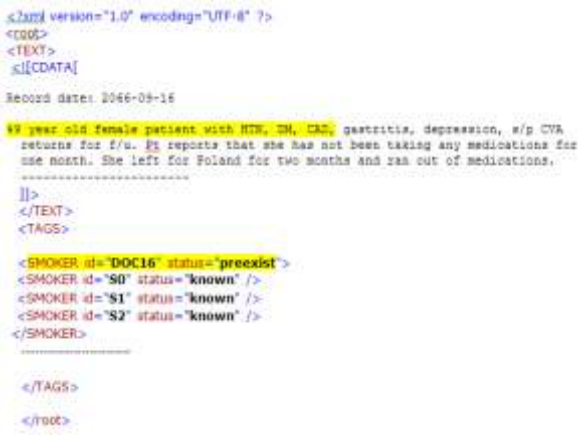

Fig. 2 Patient Data Structure

so all required paramets extracted used for predicting the risk score.Another method Reynolds risk score can be determined using a computational formula for both men and women. A 10-year cardiovascular disease for men can be estimated using equation (1) and equation (2) used for evaluating the risk for women.

$$Reynolds\ CVD\ risk = [1 - 0.8990\ (exp[B - 33.097])] \times 100\ \% \quad (1)$$

Where B = 4.385× ln (age) + 2.607× ln (BP) + 0.963×ln (Total cholesterol) – 0.772 × ln (HDL) + 0.405 (if current smoker) + 0.102 ×ln (HSCRP) + 0.541 (Parental History).

$$Reynolds\ Risk\ for\ Women = [1 - 0.98364\ (exp[B - 22.325])] \times 100\ \% \quad (2)$$

Where B = 0.0799× (age) + 3.137× ln (BP) + 1.382×ln (Total cholesterol) – 1.172 × ln (HDL) + 0.818 (if current smoker) + 0.180 ×ln (HSCRP) + 0.438 (Parental History).

Another simple method for calculating risk is 10-year prospective cardiovascular munster (PROCAM) study based on age, blood pressure, LDL cholesterol and HDL cholesterol and triglycerides. All scores are categorized into three groups , low if it is <10%, moderate if risk score is in 10-20% range, and high if the risk score is >20%.

III. SPARK PYTHON(PySpark)

Apache Spark is an open source big data processing framework built around speed, ease of use, and sophisticated analytics. Spark has several advantages compared to other big data and Map Reduce technologies like Hadoop and Storm. First of all, Spark gives us a comprehensive, unified framework to manage big data processing requirements with a variety of data sets that are diverse in nature (text data)

Spark enables applications in Hadoop clusters to run up to 100 times faster in memory and 10 times faster even when

running on disk. It comes with a built-in set of over 80 high-level operators. And we can use it interactively to query data within the shell .In addition to Map and reduce operation. It supports SQL queries, streaming data, machine learning and graph data processing. Developers can use these capabilities stand-alone or combine them to run in a single data pipeline use case. Spark allows programmers to develop complex, multi-step data pipelines using directed acyclic graph (DAG) pattern. It also supports in-memory data sharing across DAGs, so that different jobs can work with the same data. Spark runs on top of existing Hadoop Distributed File System (HDFS) infrastructure to provide enhanced and additional functionality. Spark holds intermediate results in memory rather than writing them to disk which is very useful especially when you need to work on the same dataset multiple times. It's designed to be an execution engine that works both in-memory and on-disk. Spark operators perform external operations when data does not fit in memory. Spark can be used for processing datasets that larger than the aggregate memory in a cluster. Spark will attempt to store as much as data in memory and then will spill to disk. Spark is written in scala programming language and also offers several interactive APIs such as java, python and R. the reason for selecting the python which supports Natural language processing tool kit. Figure 3 shows block diagram of risk prediction classification and analysis using PySpark and Natural Processing Tool Kit (NLTK).

Python is a high-level, interpreted, interactive and object-oriented scripting language. Python is designed to be highly readable. It uses English keywords frequently where as other languages use punctuation, and it has fewer syntactical constructions than other languages. It is open source language available with spark is known as PySpark.

NLTK is a leading platform for building Python programs to work with human language data. NLTK will aid with everything from splitting sentences from paragraphs, splitting up words, recognizing the part of speech of those words, highlighting the main subjects, and then even with helping machine to understand what the text is all about. It supports automatic methods to find characteristic words and expressions of a text. NLTK helps to identify and extract defined words and also use regular expressions to search, match and select the specific text present in unstructured data. Natural Language Processing with Python provides a practical introduction to programming for language processing. Written by the creators of NLTK, it guides the reader through the fundamentals of writing Python programs, working with corpora, categorizing text, analyzing linguistic structure, and more. So merits of PySprk with NLTK allow developing a new method for determining the risk factor.

PySpark which uses data present in Hadoop Distributed File System (HDFS) and extracts required parameters using Natural Language Processing Libraries. CAD risk level estimated for each parameter using risk procedure. Finally risk factor can be determined by averaging the all risk levels and it is expressed in terms of percentage. Further it can be classified and analysed.

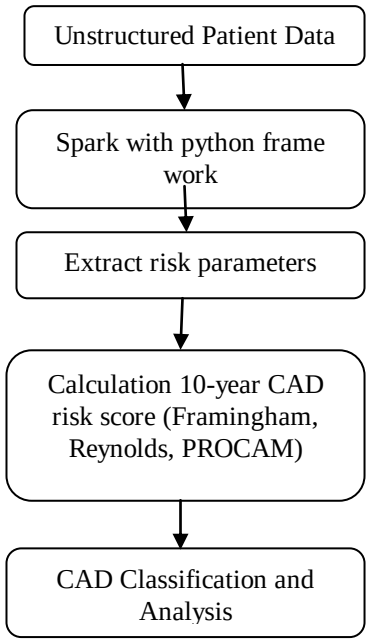

Fig. 3 Process of CAD Risk Prediction

IV. RESULTS AND DICUSSIONS

All risk prediction methods have been executed on windows based python and spark with python (PySpark), table 1 shows execution time required for different algorithms on different framework. We consider 53 data files, in which 23 files are already having CAD relevant diseases, 21 data files are having normal clinical parameters and remaining files contain missing parameters. Risk prediction methods implemented simultaneously on all files. Method in turn produces low or high risk score. If data file contains abnormal symptoms and method predicts High risk score indicates correctly predicted the risk otherwise wrongly predicted. For normal symptoms, if it detects low risk means correctly identified and high risk score indicates wrong prediction. Table 2 shows risk prediction of different methods on both normal and diseased data files.

Table 1. Execution time on different frame work

Type of Data	Frame work	Method	Execution Time (seconds)	Speed Improvement (average)
xml	Spark with python (PySpark)	Framingham prediction	0.067278	96.09%
		Reynolds Risk	0.067283	
		PROCAM	0.067266	
	Python	Framingham prediction	1.730	
		Reynolds Risk	1.719	
		PROCAM	1.716	

Table 2. True rate and false rate

	True Positive	False Negative	True Negative	False Positive
Framingham Risk	15	8	20	4
Reynolds Risk	19	4	21	3
PROCAM	17	6	15	9

Cardiac risk methods have been implemented on both normal and abnormal data. 23 data files are having abnormal files means those are having past CAD problems and symptoms, 24 are normal files and 6 are having missing data. For total 23 data files out of 53,Framingham risk prediction showing 15 data files are having high risk(disease person and symptoms shows CAD, identified as sick(high risk) and 8 are identified as low risk (normal), where as Reynolds risk score 19 are predicted as high risk and 4 files predicted as low risk. PROCAM predicts 17 are high risk and 6 are identified as normal or low risk score.

For total 24 normal data files ,Framingham risk prediction showing 20 data files are having low risk(normal person and symptoms shows also normal, identified as normal(low risk) and 4 identified as high risk (normal), where as Reynolds risk score 21 are predicted as low risk and 3 files predicted as high risk. PROCAM predicts 15 are low risk and 9 are identified as abnormal or high risk score. Finally true rate and false rate of all prediction methods have been analysed. Sensitivity and specificity are determined as follows.

$$\text{Sensitivity} = TP / (TP + FN) \quad (3)$$

$$\text{Specificity} = TN / (TN + FP) \quad (4)$$

$$\text{Accuracy} = TP + TN / TP + FN + TN + FP \quad (5)$$

Where TP = True Positive
FN = False Negative
TN = True Negative
FP = False Positive

Table 3 shows sensitivity and specificity of all prediction methods. Reynolds shows high sensitivity of 82 % and high specificity of 87%.compared to other risk prediction methods. Framingham shows sensitivity of 65% and specificity of 83%.PROCAM measures sensitivity of 73% and specificity of 62%.accuracy of Reynolds, Framingham and PROCAM are 85%, 74% and 68% respectively. So it was observed that Reynolds risk prediction method gives better screening tool for CAD risk prediction.so that CAD disease can be controlled and prevented.

Table 3. Analysis of different methods

Method	Sensitivity	Specificity	Accuracy
Framingham Risk	0.65	0.83	0.74
Reynolds Risk	**0.82**	**0.87**	**0.85**
PROCAM Risk	0.73	0.62	0.65

V. CONCLUSIONS

The paper discussed about CAD prediction and analysis of using unstructured data gives useful information for the patients about diseases progress and control. We observe Reynolds risk score is one of the best method, which determines the risk of both men and women with high sensitivity and specificity compared to other methods. We also find missing parameters in some of the data files is one of the hurdles in this method. It may be rectified using regression technique. Our work limited to few data files due to unavailability. Proposed method helps academicians, Industry and researchers the use of advanced technology in health care analytics.

REFRENCES

[1] Mozaffarian D, Benjamin EJ et al.," Heart disease and stroke statistics-update" a report from the American Heart Association, pp.4-9, 2016.

[2] D. Lloyd-Jones et al., "Heart disease and stroke statistics–2009 update" a report from the American Heart Association Statistics Committee and Stroke Statistics Subcommittee, Circulation pp.31-35, 2009.

[3] Jitendra Jonnagaddala et al., "Coronary artery disease risk assessment from unstructured electronic health records using text mining" Journal of Biomedical Informatics, volume 58,pp.203-210,2015.

[4] Hui Yang and Jonathan M. Garibaldi, "A hybrid model for automatic identification of risk factors for heart disease" Journal of Biomedical Informatics,volume 58, pp.171-182,2015.

[5] Huan Li, Kejie Lu, and Shicong Meng, "BigProvision: *A* Provisioning Framework for Big Data Analytics" IEEE Network, volume 29, Issue 5, September/October 2015.

[6] Saravana kumar N, T.Eswari, P.Sampath and S.Lavanya, **"**Predictive Methodology for Diabetic Data Analysis in Big Data", 2nd International Symposium on Big Data and Cloud Computing (ISBCC'15), Procedia Computer Science, volume 50, pp.203-208, 2015.

[7] Jimeng Sun, Jianying Hu, Dijun Luo and walter stewarts,"Combining Knowledge and Data Driven Insights for Identifying Risk Factors using Electronic Health Records" AMIA Symposium; American Medical Informatics Association,pp.90-910,2012

[8] Yue Wang, Jin Luo, NLP based congestive heart failure case finding: A prospective analysis on statewide electronic medical records, International Journal of Medical Informatics, volume 84, Issue 12,pp.1039-1047,2015.

[9] Qingcai Chena et al.." An automatic system to identify heart disease risk factors in clinical texts over time", Journal of Biomedical Informatics, volume 58, pp. 158-163, 2015.

[10] J.Archenaa and E.A.Mary Anita," A Survey Of Big Data Analytics in Healthcare and Government", 2nd International Symposium on Big Data and Cloud Computing (ISBCC'15), Procedia Computer Science, volume 50, pp.408 – 413, 2015.

[11] Nancy R. Cook et al.,"Comparison of the Framingham and Reynolds Risk Scores for Global Cardiovascular Risk Prediction in the Multiethnic Women's Health Initiative", volume 125, pp. 1748–1756, 2012.

[12] Eeva Ketola, Tiina Laatikainen and Erkki Vartiainen," Evaluating risk for cardiovascular diseases-vain or value? How do different cardiovascular risk scores act in real life", European Journal of Public Health, Volume 20, pp. 107–112,2009.

[13] Sharmini Selvarajah and Gurpreet Kaur , "Comparison of the Framingham Risk Score, SCORE and WHO/ISH cardiovascular risk prediction models in an Asian population", International Journal of Cardiology , volume 176, pp.211-218,2014.

[14] Paul M Ridker, Nina P. Paynter, Nader Rifai, J Michael Gaziano, and Nancy R Cook,"C-Reactive Protein and Parental History Improve Global Cardiovascular Risk Prediction: The Reynolds Risk Score for Men, Circulation",Pubmed, volume 118, pp. 2243–2251, 2008.

[15] https://www.i2b2.org/NLP/DataSets

[16] https://www.python.org/doc/

[17] http://www.nltk.org/install.html

BIOGRAPHY

G.Tirupati, is working as an Assistant professor in Department of Information Technology at GVP College of Engineering for Women, Visakhapatnam, India. He completed Master Degree in Biomedical Engineering and pursing M.Tech in Computer Science & Technology from Andhra University. His research interests are Data Engineering and Medical Informatics.

Dr. K.Venkata Rao, is working as a Professor in Department of Computer Science and Systems Engineering at Andhra University College of Engineering (A), Visakhapatnam. He published many international and national papers in the field of Image analysis using fuzzy algorithms and Data Structures. His research interests are Image Processing, Data Structures, Systems Programming, Programming Languages, Neural Networks, Fuzzy Systems and Web Technologies.