glucose levels known as metabolic syndrome is highly prevalent in Asian Indians.

- **Urbanisation**—The developing countries like India are undergoing rapid urbanisation Urbanization is associated with increasing obesity, decreasing physical activity due to changes in lifestyle, diet and a change from manual work to less physical occupations.
- **Stress**—The impact of stress both physical and mental along with lifestyle changes has a strong effect of increasing incidence of type II Diabetes amongst persons is a strong genetic background.

2. Literature Survey

Management of diabetes represents an enormous challenge for health systems at every level of development. The researchers are ongoing with their work continuously to deliver high quality care to patients from the day they are diagnosed throughout their life so that to control the mellitus disease. In this study, we review the status of diabetes management in our country and and try to identify the key challenges the country needs to address to reduce the present and future medical and economic burden caused by the disease.Table 1 shows a sample of different data mining techniques used in the diagnosis of Diabetes disease over different Diabetes disease datasets. The results of the different data mining research cannot be compared because they have used different datasets.

Table 1: A sample of data mining techniques used on different diabetes disease datasets

Author	*Year*	*Technique Used*	*Sample size*
Yan, et al.	2003	multilayer perception	800
Andreeva,p.	2006	Naïve bayes	Mass
		Decision tree	
		Neural n/w	
		Kernal density	
Palaniappan,et al.	2007	Naïve bayes	Mass
		Decesion tree	
		Neural n/w	
De Beule, et al.	2007	Artificial neural Network	200
Tantimongcolwata, et al.	2008	Direct kernel self organizing Map	350
		Multilayer Perceptron	200
Hara, et al.	2008	Automatically Defined(Tanagra)	400
		Immune Multi-agent Neural Network	
Sitar-Taut, et al.	2009	Naïve Byes	Mass
		Decision tree	
AJ kumar, et al.	2010	Naive Bayes(weka),KNN	500

In the above table, the various researchers Andreeva p.,Palaniappan et al. and Sitar-Taut et al used naïve bayes, neural network and decision tree data mining techniques on mass of sample size with various data mining tools. Yan et al(2003),used multilayer perception on a sample size of 800 people for diabetes prediction. De beule et al. works on artificial tech on a sample size of 200 people. Tantimongcolwata et al used direct kernel self organizing map and multilayer perception on sample size of 350 and 200 people respectively. Hara et al.(2008) using Tanagra done automatically and immune multi-agent neural network on 400 persons. AJ kumar(2010) using weka a data mining tool done work on 500 people using naïve bayes and KNN techniques.

2.1. Data Mining

Data Mining represents a process developed to examine large amounts of data routinely collected. The term also refers to a collection of tools used to perform the process. Data collected from various areas such as marketing, health, communication, etc., are used in data mining. Data Mining is the extraction of hidden predictive information from large databases; it is a powerful technology with great potential to help organisations focus on the most important information in their data warehouse. Questions those traditionally were too time consuming to resolve can be answered by the data mining tools in an effective manner. This helps to find the hidden patterns, predictive information that facilitates the experts with solution outside their expectations .The goal of data mining is to extract knowledge from dataset in human-understandable structures. In recent years data mining has been used widely in the areas of science and engineering, such as bioinformatics, genetics, medicine, education and engineering.

2.2. Techniques used in data mining

2.2.1.Clustering is the task of assigning a set of objects into groups (called clusters) so that the objects in the same cluster are more similar (in some sense or another) to each other than to those in other clusters. Clustering is a main task of explorative data mining, and a common technique for statistical data analysis used in many fields, including machine learning, pattern recognition, image analysis, information retrieval and bioinformatics.

2.2.2. Classification trees are used to predict the classes of a categorical dependent variable from their measurements on one or more predictor or independent variables. Decision Trees have emerged as a powerful technique for modeling general input / output relationships. They are tree – shaped structures that represents a series of roles that lead to sets of decisions. They generate rules for the classification of a dataset and a logical model represented as a binary (two – way split) tree that shows how the value of a target variable can be predicted by using the values of a set predictor variables. Decision trees, which are considered in a regression analysis problem, are called regression trees. Thus, the decision tree represents a logic model of regularities of the researched phenomenon.

2.2.3. Part is a rule based algorithm and produces a set of if-then rules that can be used to classify data. It is a modification of C4.5 and RIPPER algorithms and draws strategies from both. PART adopts the divide-and-conquer strategy of RIPPER and combines it with the decision tree approach of C4.5. PART generates a set of rules according to

the divide-and conquer strategy, removes all instances from the training collection that are covered by this rule and proceeds recursively until no instance remains. To generate a single rule, PART builds a partial decision tree for the current set of instances and chooses the leaf with the largest coverage as the new rule. It is different from C4.5 because the trees built for each rules are partial, based on the remaining set of examples and not complete as in case of C4.5.

2.2.4. ID3 Algorithm starts with all the training samples at the root node of the tree. An attribute is selected to partition these samples. For each value of the attribute a branch is created, and the corresponding subset of samples that have the attribute value specified by the branch is moved to the newly created child node. The algorithm is applied recursively to each child node until all samples at a node are of one class. Every path to the leaf in the decision tree represents a classification rule.

Figure 1: Data mining process

2.3. Comparative statement

The following table presents the comparative statement of various data mining trends from past to the future.

Table 2: Current data mining areas and techniques to mine the various data formats

Data mining Trends	*Algorithms/ Techniques employed*	*Data formats*	*Computing Resources*	*Prime areas of applications*
Past	Statistical, Machine Learning Techniques	Numerical and Structured data stored in traditional databases	Evolution of 4GPL and various related techniques	Business
Present	Statistical, Machine Learning, Artificial Intelligence, Pattern Reorganization Techniques	Heterogeneous data formats (structured, semi structured And unstructured Data)	High speed networks, High end storage devices and Parallel, Distributed computing etc…	Business, Web, Medical diagnosis etc…
Future	Soft Computing techniques like Fuzzy logic, Neural Networks and Genetic Programming	Complex data objects (high dimensional, high speed data streams, sequence, series, graph, Multi instance objects, Multi represented objects and temporal data etc)	Multi-agent technologies and Cloud Computing	Business, Web, Medical diagnosis, Scientific and Research analysis fields (bio, remote sensing etc…), Social networking etc…

Advantages of Using Data Mining In Various Applications Such As:

- **Banking**: Data mining supports banking sector in the process of searching a large database to discover previously unknown patterns; automate the process of finding predictive information. Data mining helps to forecast levels of bad loans and fraudulent credit cards use, predicting credit card spending by new customers and predicting the kinds of customer best respond to new loan offered by the backs.
- **Manufacturing and production**: Data mining helps to predict the machine failures and finding key factors that control optimization of manufacturing capacity.
- **Marketing:** Data mining facilitates marketing sector by classifying customer demographic that can be used to predict which customer will respond to a mailing or buy a particular product and it is very much helpful in growth of business.
- **Healthcare**: Data mining supports a lot in health care sector. It supports health care sector by correlating demographics of patients with critical illnesses, developing better insights on symptoms and their causes and learning how to provide proper treatments.
- **Insurance:** Data mining assist insurance sector in predicting fraudulent claims and medical coverage cost, classifying the important factors that affect medical coverage and predicting the customers' pattern which customer will buy new policies.
- **Law:** Law enforcement is helped by data mining by monitoring the behaviour patterns of the criminals. Tracking crime pattern, locations and criminal behaviours, identifying various attributes to data mining, assist in solving criminal cases.
- **Government and Defence:** Data mining helps to forecast the cost of moving military equipment and predicting resource consumption. Apart from that it assists in testing strategies for potential military engagements and improving homeland security by mining data from many sources.
- **Computer Hardware and Software:** Predicting disk-failures and potential security violations can be done by data mining.

Table 3: Summary of different data mining tools used on different disease predictions with accuracy

Author	*Technique used*	*Data mining tool*	*Accuracy*	*Objective*
Cheung(2001)	decision tree,naïve bayes	orange	81.11%, 81.48%	decrease the rate of heart disease
Tu et al(2009)	J4.8,decision tree,bagging algorithm	orange	78.9%,81.49%	T2DM estimation
Polat et al.(2007)	K-nearest neighbor	Weka	87.00%	lungs disease prediction
Kr laxmi,M Verra Krishna	association rule	R tool	80.03%	controlling asthma disease
Abdi et al(2013)	SVM,AR-M2P	Weka	78.11%	forcasting of cancer disease
Yan et al.	Kdd	Weka	89.03%	Development of a decision support system for heart disease

As shown in table 3, Cheung (2001) using orange tool and the techniques decision tree and naïve bayes, he tries to decrease the rate of disease, Tu et al(2009)by using same tool with different techniques(j4.8 and bagging algorithm) also tries to reduce chronic disease, Polat et al.(2007) predict lungs disease by using weka tool, Das et al(2009) used Tanagra tool and by using neural network technique shows accuracy of prediction 89.01%. Svijyarani, S sudha(2013), Jan G Bazan, Stannislawn(2009) are also work on to control the mellitus disease by using weka tool, Sunita Behar, Mr Lobo Mrj(2011) using Tanagra using clustering technique to describe the stage of disease using kdd method, using R tool, Kr laxmi, M Verra Krishna, controlling asthma disease with accuracy 80.03%, Abdi et al(2013) forcast the cancer disease, Van et al. using kdd method develop a decision support system for heart attack.

2.4. Open source tools for data mining

Different types of data mining tools are available in software market, each with its own strengths and weaknesses. Some of the data mining tools available today are explained as follows:

2.4.1. R Tool is an open source programming language and environment for statistical computing and graphics. R provides a wide variety of graphical and statistical techniques such as linear and non-linear modelling, classical statistical tests, time series analysis, classification clustering and is highly extensible. Researchers in various fields of applied statistics have adopted R for statistical software development and data analysis. Extensibility and superb data visualisation are the two main reasons for the success of R

2.4.2. Weka is a collection of machine learning algorithms for data mining tasks and well suited for developing new machine learning schemes. Weka is a java based software capability of working under various operating systems and contains tools for data pre-processing, classification, regression, clustering, association rules and visualization. The algorithms can either be applied directly to a dataset or called from a user's java code. Weka is probably the most successful open source data mining software which has inspired by the development of other programs with more sophisticated graphical user interface and better visualization methods.

2.4.3. Orange is an open source data mining and visualisation software with active community and which helps novice and experts for their analysis. It has the ability to work under various platforms like windows, Mac Os C and GNU/Linux operating systems and it's packed with data analytics features. It enables design of data analysis process through user friendly visual programming or python scripting. Hence, this can be used as a scripting language for respective tasks of data mining. It represents most major algorithms for data mining and contains different visualisation, from scatter plots , bar charts, trees to diagrams, networks and heat maps. It remembers user's choices, suggests most frequently used combinations, and intelligently chooses which communication channels to use. It has specialised add-ons like Bio orange for bio informatics

2.4.4. Rapid Miner is an open source system for data mining which is available as a standalone application for data analysis and as a data mining engine for the integration into own products. It has ability to runs on major platform and operating systems. It is powerful but intuitive graphical user interface for the design of analysis processes. It offers data integration, analytical ETL, Data analysis and reporting in one single suite. It provides a graphical process design for standard tasks and scripting language for arbitrary operations.

2.4.5. Tanagra is open source data analysis software for academic and research purposes which proposes several data mining methods from exploratory data analysis, statistical learning, machine learning and databases area. The main purpose of Tanagra is to provide researchers and students to use data mining software in an easy way by conforming to the present norms of the software development and allowing to analyse either real or synthetic data. The second purpose is to propose an architecture allowing the users to add to add their own data mining methods it helps to compare their performances. It acts more as an experimental platform in order to do the essential work, dispensing them to deal with the unpleasant part of the data management. Last purpose is to give the direction to novice developers in diffusing a possible methodology for building this kind of software. It can be considered as a pedagogical tool for learning programming techniques since it permits to access the source code, to look pattern of the software how it is built, the problems to avoid, key steps of the project, tools used and code libraries used for the project.

Table 4 illustrates a sample of data mining techniques used in the diagnosis of Diabetes disease. Researchers have used data mining techniques on the Diabetes disease benchmark dataset to extract trends and relationships between different variables such as blood pressure and cholesterol.

Table 4: A Sample of Data used on diabetes disease dataset using different data mining techniques

Author	Year	Technique	Accuracy
Cheung	2001	Decision Tree	81.11%
		Naïve Bayes	81.48%
Tu, et al.,	2009	J4.8 Decision Tree	78.9%
		Bagging algorithm	81.41%
Polat et al.	2007	Fuzzy-AIRS–k-nearest neighbour	87%
Das, et al., 89.01%	2009	Neural network ensembles	89.01%
Kavitha et al.	2010	neuralnetwork, genetic algorithm	88.05%

Recently, researchers started using hybrid data mining techniques in the diagnosis of Diabetes disease. Polat et al. used fuzzy artificial immune recognition system and k-nearest neighbour in the detection of Diabetes disease. The proposed model showed accuracy of 87% in the detection of Diabetes disease patients. Das et al., used neural network ensembles in the diagnosis of Diabetes disease showing accuracy of 89.01%. Comparison of single and hybrid data mining techniques in the diagnosis of Diabetes disease shows different accuracies, with the hybrid techniques showing better accuracy than single techniques. The best accuracy achieved using single data mining technique is 84.14% by naïve-bayes. However, the best accuracy achieved using hybrid data mining technique is 89.01% by neural network ensemble. Hybridized data mining techniques are enhancing the accuracy of Diabetes disease diagnosis. Through a systematic investigation of several single data mining technique, different data discretization levels, the application of voting techniques, and reduced error pruning, we demonstrated increases in accuracy on all techniques assessed.

3. Methodology Used in Data Mining

Different data mining techniques have been used to help health care professionals in the diagnosis of Diabetes disease. Those most frequently used focus on classification: naïve bayes decision tree, and neural network. Other data mining techniques are also used including kernel density, automatically defined groups, bagging algorithm, and support vector machine. Though applying data mining is beneficial to healthcare, disease diagnosis, and treatment, few researches have investigated producing treatment plans for patients.

3.1. Pre-processing

Before data mining algorithms can be used, a target data set must be assembled. As data mining can only uncover patterns actually present in the data, the target data set must be large enough to contain these patterns while remaining concise enough to be mined within an acceptable time limit. A common source for data is a data mart or data warehouse. Pre-processing is essential to analyze the multivariate data sets before data mining. The target set is then cleaned. Data cleaning removes the observations containing noise and those with missing data.

3.2. Extraction of Useful Knowledge

- Brings a set of tools and techniques that can be applied to this processed data to discover hidden patterns
- That provide healthcare professionals an additional source of knowledge for making decisions
- The decisions results then matched with health care professionals opinions.

Simply stated, “Data mining refers to extracting or “mining” knowledge from large amounts of data”. There are some other terms which carry a similar or slightly different meaning to data mining, such as knowledge mining from data, knowledge extraction, data or pattern analysis, and data archaeology. Data mining functionalities are used to specify the kind of patterns to be found in data mining tasks. In general, classified data mining tasks into two categories: descriptive and predictive. Descriptive mining tasks characterize the general properties of the data in the database where as predictive mining tasks perform inference on the current data in order to make predictions. Data mining, also popularly known as Knowledge Discovery in Database refers to the process of discovering interesting knowledge from large amounts of data stored in databases, data warehouses, or other information repositories. Data mining techniques are used to operate on large volumes of data to discover hidden patterns and relationships helpful in decision making. While data mining and knowledge discovery in database are frequently treated as synonyms, data mining is actually part of the knowledge discovery process.

3.3. Result Validation

Data mining can unintentionally be misused, and can then produce results which appear to be significant; but which do not actually predict future behavior and cannot be reproduced on a new sample of data and bear little use. Often this results from investigating too many hypotheses and not performing proper statistical hypothesis testing. A simple version of this problem in machine learning is known as over fitting, but the same problem can arise at different phases of the process and thus a train/test split - when applicable at all - may not be sufficient to prevent this from happening The final step of knowledge discovery from data is to verify that the patterns produced by the data mining algorithms occur in the wider data set. Not all patterns found by the data mining algorithms are necessarily valid. It is common for the data mining algorithms to find patterns in the training set which are not present in the general data set. This is called over fitting. To overcome this, the evaluation uses a test set of data on which the data mining algorithm was not trained. The learned patterns are applied to this test set, and the resulting output is compared to the desired output. For example, a data mining algorithm trying to distinguish "spam" from "legitimate" emails would be trained on a training set of sample e-mails. Once trained, the learned patterns would be applied to the test set of e-mails on which it had not been trained. The accuracy of the patterns can then be measured from how many e-mails they correctly classify.

4. Conclusion

Motivated by the world-wide increasing mortality of diabetes disease patients each year and the availability of huge amounts of data, researchers are using data mining techniques in the diagnosis of diabetes disease. Although applying data mining techniques to help health care professionals in the diagnosis of diabetes disease is having some success, the use of data mining techniques to identify a suitable treatment for diabetes disease patients has received less attention. Also, applying data mining techniques has shown promising results in the diagnosis of diabetes disease, so applying data mining techniques in selecting the suitable treatment for diabetes disease patients needs further investigation. This paper identifies the research on diabetes disease diagnosis and treatment and proposes a model to systematically close to discover if applying data mining techniques to diabetes disease treatment data can provide as reliable performance as that achieved in diagnosing diabetes disease patients.

References

[1] Ashby, D. and A. Smith, The Best Medicine Plus Magazine-Living Mathematics, 2005.
[2] AJ kumar, et al.(2005),using KNN and naive bayes technique to on weka-a data mining tool.
[3] Abdi et al(2013) using weka forcast the cancer disease using svm,ar-m2p technique.
[4] Andreeva, P., Data Modelling and Specific Rule Generation via Data Mining Techniques. International Conference on Computer Systems and Technologies - CompSysTech, 2006.
[5] Das, R., I. Turkoglu, and A. Sengur, Effective diagnosis of heartdisease through neural networks ensembles. Expert Systems with Applications, Elsevier, 2009. 36 (2009); p. 7675–7680.
[6] European Public Health Alliance. 2010 7-February-2011]; Available from: http://www.epha.org/a/2352 176 2012 Japan-Egypt Conference on Electronics, Communications and Computers.
[7] Han, J. and M. Kamber, Data Mining Concepts and Techniques.2006: Morgan Kaufmann Publishers.
[8] Helma, C., E. Gottmann, and S. Kramer, Knowledge discovery and data mining in toxicology. Statistical Methods in Medical Research, 2000.
[9] Heller, R.F., et al., How well can we predict coronary heart disease? Findings in the United Kingdom Heart Disease Prevention Project. BRITISH MEDICAL JOURNAL, 1984.
[10] Li L, T.H., Wu Z, Gong J, Gruidl M, Zou J, Tockman M, Clark RA, Data mining techniques for cancer detection using serum proteomic profiling. Artificial Intelligence in Medicine, Elsevier, 2004.
[11] Lee, I.-N., S.-C. Liao, and M. Embrechts, Data mining techniques applied to medical information. Med. inform, 2000.
[12] Liao, S.-C. and I.-N. Lee, Appropriate medical data categorization for data mining classification techniques. MED. INFORM., 2002.Vol. 27, no. 1, 59–67.
[13] Obenshain, M.K., Application of Data Mining Techniques to Healthcare Data. Infection Control and Hospital Epidemiology,2004.
[14] Podgorelec, V., et al., Decision Trees: An Overview and Their Use in Medicine. Journal of Medical Systems, 2002. Vol. 26.
[15] Porter, T. and B. Green, Identifying Diabetic Patients: A Data Mining Approach. Americas Conference on Information Systems,2009.
[16] Panzarasa, S., et al., Data mining techniques for analyzing stroke care processes. Proceedings of the 13th World Congress on Medical Informatics, 2010.
[17] Rajkumar, A. and G.S. Reena, Diagnosis Of Heart Disease Using Datamining Algorithm. Global Journal of Computer Science and Technology, 2010. Vol. 10 (Issue 10).
[18] Ruben, D.C.J., Data Mining in Healthcare: Current Applications and Issues. 2009.

Author Profile

Ravneet Jyot Singh received the B-tech. degrees in Computer Science Engineering in 2010 Punjab Technical University, Jalandhar (Punjab) and now he is pursuing M-tech. in Computer Engg. From Punjabi University Patiala(Punjab).

Williamjeet Singh received his B.Tech. and M.Tech degree in Computer Science and Engineering in 2005 and 2007 respectively from Punjab Technical University, Jalandhar (Punjab) and Punjabi University, Patiala (Punjab). Presently he is working as Assistant Professor and pursuing his PhD from Computer Engineering Department, Punjabi University, Patiala. His current research interests are primarily in the area of Wireless networks, Algorithms and Data mining.