OPEN ACCESS

Middle East Research Journal of Engineering and Technology
ISSN **2789-7737 (Print) & 2958-2059 (Online)**
Frequency: Bi-Monthly
DOI: 10.36348/merjet.2021.v01i01.007

KUWAIT SCHOLARS PUBLISHER
Website: http://www.kspublisher.com/
Email: office.kspublisher@gmail.com

Analysis on Diabetes Prediction Using Different Classifiers Techniques

Amanpreet Kaur[1*], Gagandeep Singh[1], Beant Kaur[1]
[1]Assistant Professor, CSE Department of BGIET, Sangrur, Punjab, India

ABSTRACT: Huge amount of data is fetched from various resources. Data that is required from large data repository can be extracted with the use of data mining techniques, as there might be relevant and irrelevant data present in the repository. Thus minimizes the access of data from repository, so for increasing the data usage ability, data mining techniques are required. In data mining data is classified in different classes which are then used to implement a various disease prediction model. Similarly diabetes can be predicted using these prediction models based on different parameters. Parameters include Pregnancies, Blood Pressure, Skin Thickness, Glucose, Insulin, Diabetes Pedigree Function, Age and many more various classifiers. Different classifiers like CART, RF, SVM, LDA, KNN are being applied on PIMA dataset for the data at the confidence level of 0.95. CART has highest accuracy of 88%. Lowest accuracy is of KNN. So prediction using CART based classifiers is best.

Keywords: CART, Random Forest, SVM, Linear discriminate analysis, KNN.

RESEARCH PAPER

***Corresponding Author:**
Amanpreet Kaur
Assistant Professor, CSE Department of BGIET, Sangrur, Punjab, India

How to cite this paper:
Amanpreet Kaur *et al.;* "Analysis on Diabetes Prediction Using Different Classifiers Techniques". Middle East Res J. Eng. Technol, 2021 Nov-Dec 1(1): 41-46.

Article History:
| Submit: 13.10.2021 |
| Accepted: 20.11.2021 |
| Published: 30.12.2021 |

I. INTRODUCTION

In recent time various types of services are emerging in the society. These services are related to the different fields of the society. Out of those fields major field is medical. India is a large country where large population resides. Various types of organized and unorganized medical facilities are available in the country. But due to the population explosion each facility remains scarce. To overcome and catalyst the growth in this part of the applications. Various researchers are involved which are growing with different researches so that the problem of scarcity of the resources can be catered without increasing the much cost.

Big Data is one of the major thrust areas. That can solve the problem of this medical mismanagement. There will be a Big Data of the medical data. That is consisting of integration of various small city level data. Any company will provide processing ability which can process this large integration of the data. So that any patient data can be provided at the required place. His case history or we can say medical history can be recorded at each step.

These data are mainly stored and isolated in disparate local systems, and are underutilized in terms of data analysis and knowledge discovery. Data mining techniques have made computational infrastructure capable of handling such enormous information burst in a cost-effective way. Up to the present, most works focus on migrating healthcare IT system and data storage to the cloud platform rather than taking advantage of information hidden in the data. A typical healthcare Data Mining system has a hierarchical structure including system layer, control layer, and service layer. System layer constructs fundamental storage and computing environment using distributed computing resources, storage resources, and network resources.

In control layer, system administrators control the load balancing, monitor system performance, and build programming environment. Finally service layer is responsible for providing large-scale healthcare services via real time management, privacy protection, and data analysis.

1.2 Types of Diabetes

Type 1 Diabetes is called insulin-dependent diabetes mellitus (IDDM) or juvenile-onset diabetes. Autoimmune, genetic, and environmental factors are involved in the development of this type of diabetes.

Type 2 Diabetes is called non-insulin-dependent diabetes mellitus (NIDDM) or adult-onset diabetes. In the type 2 diabetes, the pancreas usually

produces some insulin the amount produced is not enough for the body's needs, or the body's cells are resistant to it.

1.3 Big Data

The amount of data being generated inside and outside each enterprise has exploded. The increasing volume and detail of information, the rise of multimedia and social media, and the Internet of Things are expected to fuel continued exponential data growth for the foreseeable future.

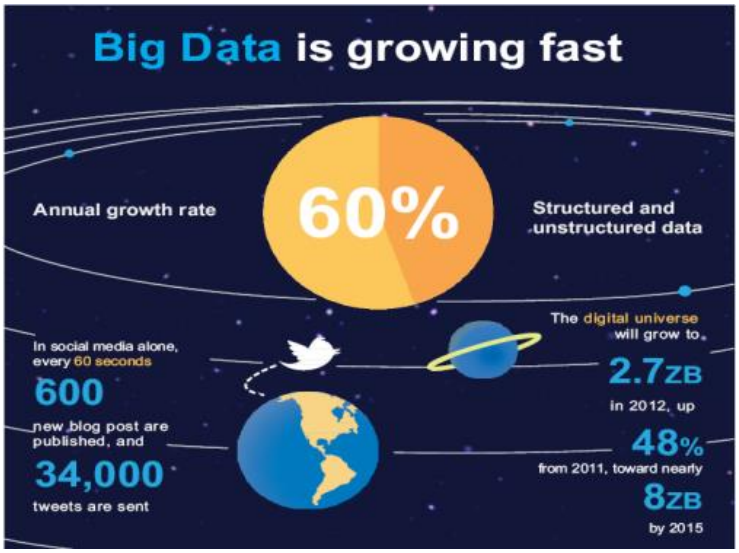


Fig. 1: Big Data Structure [2]

Two common sources of big data exist. First, there is organizational data, which, thanks to improved automation and broader access, is increasingly being shared. Organizational data includes emails, system logs, internal documents, business process events, and other structured, unstructured, and semi-structured data. It also includes social content, such as any blogs and wikis that are available within the organization.

Second, data comes from sources outside of the organization. Some of this extra-organizational data is available publicly and at no charge, some of it is obtained through paid subscription, and the rest is selectively made available by specific business partners or customers. This includes information from social media sites and various other sources, including product literature; corporate information; health, life, and consumer websites; helpful hints from third parties, and customer complaints, such as complaints posted on the websites of regulatory agencies.

1.3 Big Data Challenges in Healthcare

Inferring knowledge from complex heterogeneous patient sources. Leveraging the patient/data correlations in longitudinal records. Understanding unstructured clinical notes in the right context.

- Efficiently handling large volumes of medical imaging data and extracting potentially useful information and biomarkers.
- Analyzing genomic data is a computationally intensive task and combining with standard clinical data adds additional layers of complexity.
- Capturing the patient's behavioral data through several sensors; their various social interactions and communications.
- Take advantage of the massive amounts of data and provide right intervention to the right patient at the right time.
- Personalized care to the patient
- Potentially benefit all the components of a healthcare system Personalized care to the patient. i.e., provider, payer, patient, and management.

2. TECHNIQUES

2.1 Cart

It is classification and regression tree for the machine learning based technique where various types of data items are being classified based on tree based technique. This type of decision tree based technique can be useful for the prediction based modeling technique. This decision tree represents the root element as the single input variable and a split point on that variable. The leaf node of the tree contains the output variable like yes and no. Further another parent node represents the variable then further the grant children having yes or no.

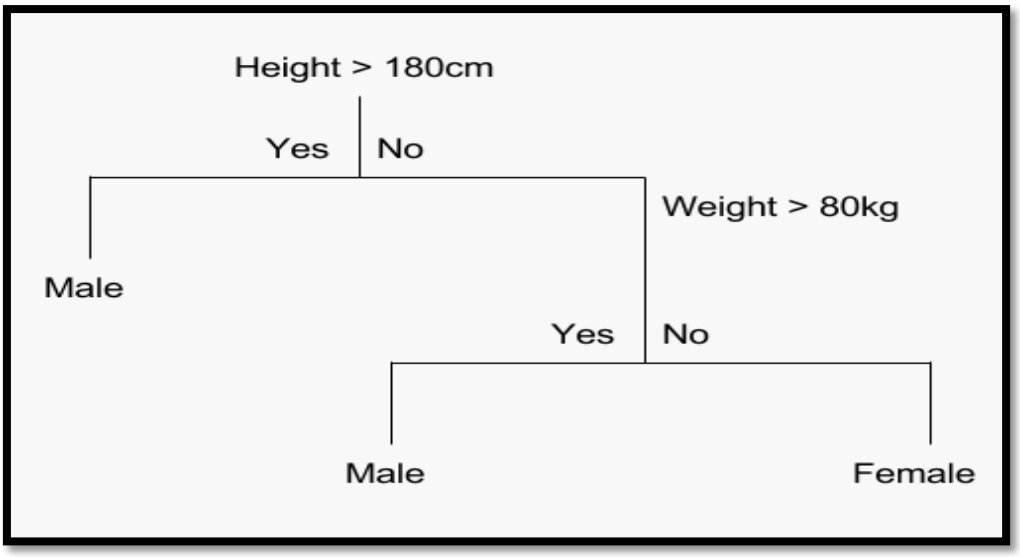


Fig. 2: CART based tree

The values of the tree are stored into the file. This stored file will be having various rules making the system of prediction quite easy. For a new input the tree will be traversed from the root and move in downward direction. Based on the input appropriate decision can be derived.

2.2 Random Forest

Random forest is another classifier used for the prediction. Currently it is the efficient technique for classification of the diabetic data. It includes various small decision trees. Each decision tree helps in having decision for the specific classification. Higher is the number of tree higher will be the accuracy. Decision tree concept is more rule based system. Given a training dataset the decision tree algorithm will be used for deriving various trees. Once the decision tree is being established, the rules are applied on to the testing set. Suppose if the person want to predict whether person will like the movie or not. For the prediction purpose various types of decision tree will be required. Each decision tree will be having different types of lead actor in the film. Now for the person who has seen the movie and like the most has certain lead actor. Based on the decision tree if this movie has the same lead actor then the person definitely will like the movie else will not like the movie.

2.3 Linear Discriminant Analysis

It is the supervised learning technique. It simply classifies the whole dataset into two or more classes. The resulting combination may be used for linear classifier or may for the dimensionality reduction. LDA is closely related is ANOVA (analysis of variance). Which also attempt to express one dependent variable as a linear combination of other features or measurements. LDA is also closely related to principal component analysis (PCA) and factor analysis in that they both look for linear combinations of variables which best explain the data. LDA explicitly attempts to model the difference between the classes of data. PCA on the other hand does not take into account any difference in class, and factor analysis builds the feature combinations based on differences rather than similarities.

2.4 KNN

KNN assumes that the data is in a feature space. More exactly, the data points are in a metric space. The data can be scalars or possibly even multidimensional vectors. Since the points are in feature space, they have a notion of distance – This need not necessarily be Euclidean distance although it is the one commonly used. Each of the training data consists of a set of vectors and class label associated with each vector. In the simplest case, it will be either + or – (for positive or negative classes). But KNN, can work equally well with arbitrary number of classes. We are also given a single number "k". This number decides how many neighbors (where neighbors are defined based on the distance metric) influence the classification. This is usually a odd number if the number of classes is 2. If k=1, then the algorithm is simply called the nearest neighbor algorithm.

3. Datasets

Kappa						
Min.	**1st Qu.**	**Median**	**Mean**	**3rd Qu.**	**Max.**	**NA's**
CART 0.15384615	0.456294	0.514923	0.536164	0.647947	0.930233	0
LDA 0.05990783	0.382867	0.477086	0.475126	0.565906	0.790698	0
SVM 0.11374408	0.380488	0.474672	0.477213	0.619801	0.793522	0
KNN 0.14671815	0.30273	0.441558	0.429303	0.529555	0.790698	0
RF 0.22123894	0.349121	0.511628	0.503959	0.608541	0.863636	0

Different parameters are considered for the prediction of the diabetes patients. The attributes of dataset are either related directly or indirectly. It is PIMA dataset.

Pregnancies	**Glucose**	**Blood Pressure**	**Skin Thickness**	**Insulin**	**BMI**	**Diabetes Pedigree Function**	**Age**	**Outcome**
6	148	72	35	0	33.6	0.627	50	1
1	85	66	29	0	26.6	0.351	31	0
8	183	64	0	0	23.3	0.672	32	1
1	89	66	23	94	28.1	0.167	21	0
0	137	40	35	168	43.1	2.288	33	1
5	116	74	0	0	25.6	0.201	30	0
3	78	50	32	88	31	0.248	26	1
10	115	0	0	0	35.3	0.134	29	0
2	197	70	45	543	30.5	0.158	53	1

4. COMPARISON RESULTS

Collected dataset of the patients on different health parameters named as PIMA dataset has been under gone through different classifiers.

The CART based technique performance in comparison to other techniques is better. So Cart based technique performance is comparatively efficient for the classification of the PIMA dataset into positive and negative classes.

4.1 Table for Accuracy

Table 1: Accuracy

Accuracy							
	Min.	**1st Qu.**	**Median**	**Mean**	**3rd Qu.**	**Max.**	**NA's**
CART	0.666667	0.759358	0.787879	0.805184	0.848485	0.969697	0
LDA	0.617647	0.757576	0.787879	0.782217	0.812166	0.909091	0
SVM	0.65625	0.759358	0.787879	0.790027	0.842246	0.911765	0
KNN	0.617647	0.709099	0.757576	0.758183	0.792558	0.909091	0
RF	0.65625	0.735294	0.787879	0.790116	0.842246	0.939394	0

Table represents the values for the CART, LDA, SVM KNN, and RF at the Confidence level of 0.95. CART based technique has optimal accuracy and the Random forest has least accuracy.

Table 2: Kappa Comparison

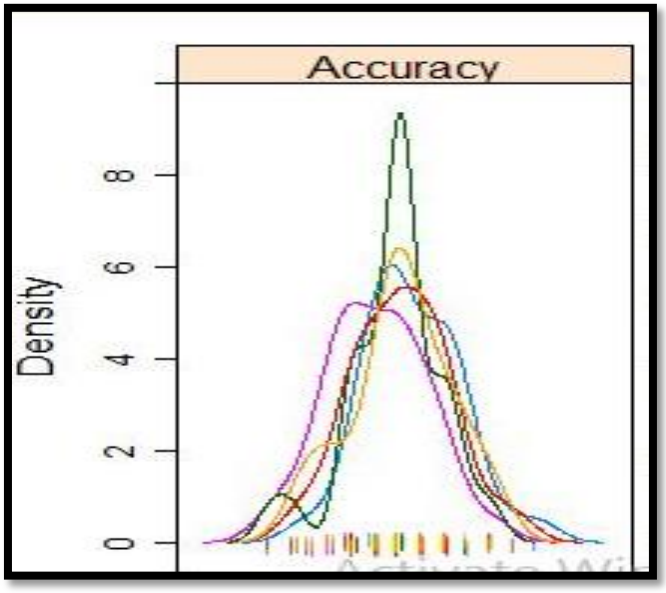
Kappa						
Min.	**1st Qu.**	**Median**	**Mean**	**3rd Qu.**	**Max.**	**NA's**
RF 0.22123894	0.349121	0.511628	0.503959	0.608541	0.863636	0
SVM 0.11374408	0.380488	0.474672	0.477213	0.619801	0.793522	0
KNN 0.14671815	0.30273	0.441558	0.429303	0.529555	0.790698	0

Cohen's kappa coefficient (κ) is a statistic which measures inter-rater agreement for qualitative (categorical) items. It is generally thought to be a more robust measure than simple percent agreement calculation, as κ takes into account the possibility of the agreement occurring by chance.

4.2 Table of Kappa

In current table the Kappa parameter is compared for all the classifiers like CART, LDA, SVM, KNN, RF. MAX kappa value is for CART based technique. The least value for the Kappa technique is for the LDA based technique.

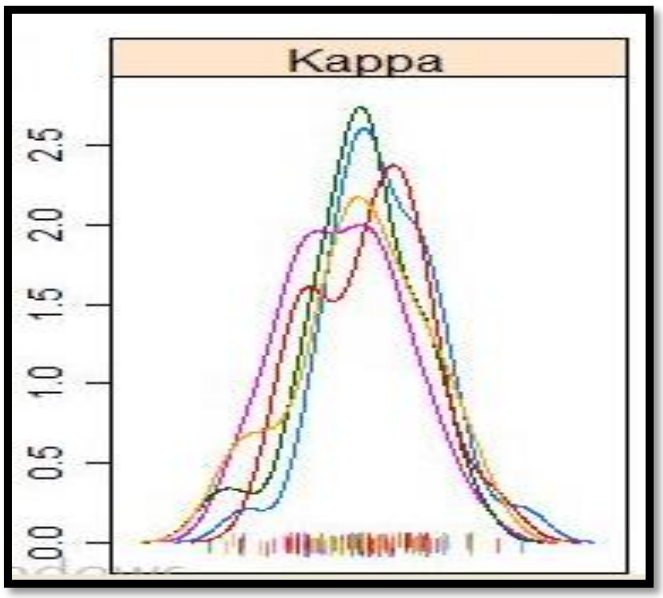
4.3 Graph for Accuracy Comparison



Graph 1: Comparative Accuracy

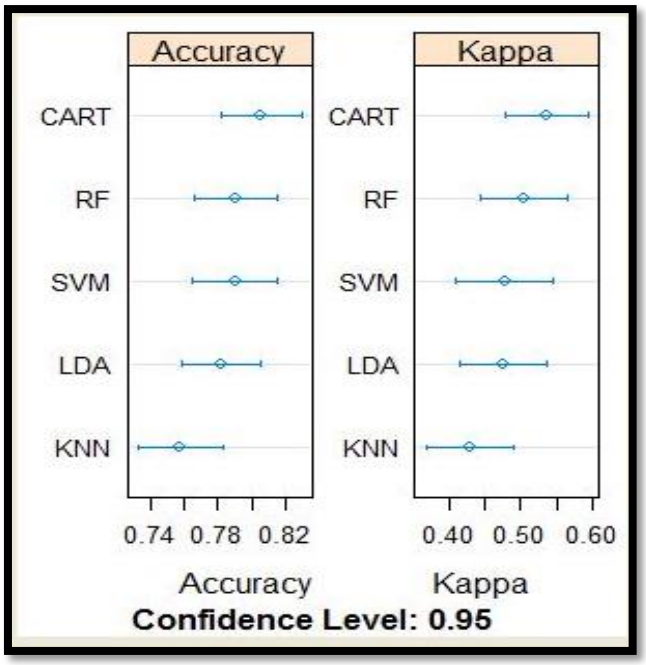
This graph shows the accuracy for the CART based technique as the maximum. And the LDA based technique has lowest accuracy.

4.4 Graph for the Kappa



Graph 2: Kappa

This graph shows the Kappa comparison for different techniques. These techniques have best technique as CART based technique. It has max Kappa value.



Graph 3: Line graph for accuracy and Kappa

This graph shows that the accuracy and Kappa value are maximum for CART based technique. The evaluation is done at the confidence value of 0.95.

5. CONCLUSION

Data mining is the most prevalent field in the area of the extraction of the relevant data. There are various applications which are generating data in large amount. These data items are in billions of bytes on every day basis. Traditional data processing applications are not enough which can process these data items on larger scale. There are different classification techniques which are used for the classification of the large dataset into two or more different classes. Later on certain prediction based system will be required for generating accurate prediction. In current research there are different classifiers are being used which has been applied on the PIMA dataset for the purpose of prediction. So that the patient whether will be having diabetes or not. Various classifiers like LDA, Random Forest, SVM, CART, KNN has been applied. The result for the accuracy and Kappa factor has been evaluated at the confidence level of 0.95. From all the classifier CART based technique has been best in both the factors like accuracy and Kappa.

6. FUTURE WORK

Mining is the most important aspect as far as system of decision making is concerned. There are various applications which are used by different businesses for the enhancement of the quality of the decision. Various classification techniques are there which can be used for the purpose of prediction. In future the accuracy can be enhanced further by hybridization of classifiers.

REFERENCES

- Aliza, A., & Aida, M. (2011). Comparison between Neural Networks against Decision Tree in Improving Prediction Accuracy for Diabetes Mellitus. *ICDIPC*, 2011, Part I, CCIS 188, pp. 537–545.
- Andreas, B., Young-Seop, L. (2001). Data Mining Criteria for Tree-Based Regression and Classifiaction. *ACM*, 1-5811-391-x/01/08
- Charu, C. Aggarwal, & Philip, S. Y. (1999). Data Mining Techniques for Associations,Clustering and Classification", c_Springer-Verlag Berlin Heidelberg, N. Zhong and L. Zhou (Eds.): PAKDD'99, LNAI 1574, pp. 13–23.
- Dursun, D., Christie, F., Charles, M., & Deepa, R. (2009). Analysis of healthcare coverage: A data mining approach. *Elsevier Expert Systems with Applications*, 36, 995–1003.
- Emoto, T., Yamashita, T., Kobayashi, T., Sasaki, N., Hirota, Y., Hayashi, T., ... & Hirata, K. I. (2017). Characterization of gut microbiota profiles in coronary artery disease patients using data mining analysis of terminal restriction fragment length polymorphism: gut microbiota could be a diagnostic marker of coronary artery disease. *Heart and vessels*, *32*(1), 39-46.
- Gao, H., Shiji, S., Jatinder, N., Gupta, D., & Cheng, W. (2014). Semi-Supervised and Unsupervised Extreme Learning Machines. *IEEE*, 2168-2267.
- Hai-Dong, M., Yu-Chen, S., Fei-Yan, S., & Hai-Tao, S. (2011). Research and application of cluster

and association analysisin geochemical data processing. *Springer Comput Geosci*, 15, 87–98.

- Humar, K., & Novruz, A. (2008). Design of a hybrid system for the diabetes and heart diseases. *Expert Systems with Applications*, 35, 82–89.
- Ioannis, K., Olga, T., Athanasios, S., Nicos, M., Ioannis, V., & Ioanna, C. (2017). Machine Learning and Data Mining Methods in Diabetes Research. Elsevier, *Computational and Structural Biotechnology Journal*, 15, 104–116.
- Ki-Hyun, K., Ehsanul, K., & ShaminAra, J. (2016). The use of cell phone and insight into its potential human health impacts, Environ Monit Assess.
- Marcano-Cedeño, A., Joaquín, T., & Diego, A. (2011). A Prediction Model to Diabetes Using Artificial Metaplasticity. IWINAC 2011, Part II, LNCS 6687, pp. 418–425.
- Rebecca, S., & Marlene, R. (2016). A user-centered model for designing consumer mobile health (mHealth) applications (apps). *Journal of Biomedical Informatics*, 60, 243–251.
- Renuka Devi, M., & Shyal. (2016). Exploring various data mining techniques, 43, 189-196.
- Rosa, M., Giuseppe, P., & Stefano, C. (1998). An Extension to SQL for Mining Association Rules. *Data Mining and Knowledge Discovery*, 2, 195–224.