https://doi.org/10.29042/2020-10-2-136-142

Diabetes Prediction using Machine Learning Techniques

[*1]Dr.O.Obulesu, [2]Dr.K.Suresh, [3]B. Venkata Ramudu

[1,3] Department of Computer Science and Engineering, Malla Reddy Engineering College (Autonomous), Secunderabad, Telagana-500100

[2]Department of Computer Science and Engineering, Sree Vidyanikethan Engineering College, Tirupati, AP-517102

Email: obulesh194@gmail.com, sureshkallam@gmail.com, bvramu@gmail.com

Received: 06th December 2019, Accepted: 20th January 2020, Published: 30th April 2020

Abstract

Diabetes Mellitus is one of the growing fatal diseases all over the world. It leads to complications that include heart disease, stroke, and nerve disease, kidney damage. So, Medical Professionals want a reliable prediction system to diagnose Diabetes. To predict the diabetes at earlier stage, different machine learning techniques are useful for examining the data from different sources and valuable knowledge is synopsized. So, mining the diabetes data in an efficient way is a crucial concern. In this project, a medical dataset has been accomplished to predict the diabetes. The R-Studio software was employed as a statistical computing tool for diagnosing diabetes. The PIMA Indian database was acquired from UCI repository will be used for analysis. The dataset was studied and analyzed to build an effective model that predicts and diagnoses the diabetes disease earlier.

Keywords

Diabetes, Classification, Clustering, Regression, SVM, K-NN, Neural Networks

Introduction

As we all know that the growth in technology helps the computers to produce huge amount of data. Additionally, such advancements and innovations in the medical database management systems generate large volumes of medical data. Healthcare industry contains very large and sensitive data. This data needs to be treated very careful to get benefitted from it. Diabetic Mellitus is a set of associated diseases in which the human body is unable to control the quantity of sugar in the blood. It results in high sugar levels in blood, may be as the physique cannot create sufficient insulin, or may because cells should not respond to the formed insulin. The focus is to develop the prediction models by using certain system trained methods. The system learning is one of the branches of synthetic intelligence as it helps the computer to learn on its own.[1] The two classification of ML are supervised and unsupervised. The Supervised learning calculation utilizes the past experience to influence expectations on new or inconspicuous information while unsupervised calculations to can draw derivations from datasets.

Machine learning algorithms are:

Supervised Learning Techniques:

Classification

This is a procedure of detecting the obscure information label name utilizing recently known (preparing information) class mark.[2] The familiar supervised learning methods are mentioned as follows.

i. Random forest
ii. SVM
iii. K-Nearest neighbors
iv. Decision tree
v. Naïve Bayes

Regression

It is also a prediction modeling method such as classification. It detects the association between the free (given) features with few target (new) features.

The popular algorithms are:

i. Simple Linear Regression
ii. Multiple Linear Regression
iii. Logistic Regression
iv. Polynomial Regression
v. Linear Discriminant Analysis (LDA)

Unsupervised Learning Techniques

Clustering

Cluster investigation or grouping is a process of classifying the similar instances into classes called groups.[3] Some of the grouping methods are as follows:

i. Mean based grouping for the given number of clusters ‘k’

ii. Level-by-level grouping based on similarity distance metrics.

R-Studio Tool

R studio 3.4.1 is used in this study. R-Studio is a United Improvement Environment for R programming language founded by JJ Allaire. It uses command line interpreter. It is used for statistical computing and graphics. Since it is having many built-in packages it can manipulate huge dataset for analysis

Literature Review

S.No.	Paper	Author(s)	Name of the Journal	Methods	Findings	Notes/Critique
1	Detecting Diabetes in Health Data by using Machine Learning Methods	U. Ali Zia, Dr. N. Khan.	International Journal of Scientific and Engineering Research (IJSER).	Boot strapping resampling technique to enhance the accuracy and then applying i. Bayes Theorem ii. Decision Trees iii. k-NN Method	After Bootstrapping Accuracy: i. NaiveBayes-74.89%; ii. Decision Trees-94.44%; iii. k-NN(for k=1) 93.79%; 4. k-NN(for k=3) -76.79%	i. Plan to use further more advanced classifiers such as Neural Networks. ii. It should consider some other important factors that are related to gestational diabetes, like metabolic syndrome, family history, habit of smoking, lazy routines, some dietary patterns etc.
2	Prediction of Diabetes Using Data Mining Techniques	Fikirte Girma, Woldemichael, Sumitra Menaria	International Conference on Trends in Electronics and Informatics (ICOEI)	i. Back Propagation Algorithm; ii. J48 Algorithm; iii. Naïve Bayes Classifier; iv. Support Vector Machine.	Back Propagation Algorithm has Accuracy-83.11% ; Sensitivity-86.53%; Specificity-76%	i. Increment the accuracy of the algorithms.
3	Diabetes Illness Detection Using Knowledge Mining Methods	D. Shetty, K. Rit, S. Shaikh, N. Patils	International Conference on Innovations in Information, Embedded and Communication Systems(ICIIECS).	i. Bayes Theorem; ii. k-NN algorithms	Detection of the illness can be finished by using Bayes theorem and K-NN method and analyze them based on various attributes of diabetes.	i. Increment the accuracy of the algorithms.; ii. Extending analysis on much attributes so to handle diabetes effectively.
4	Effective Detection & Classification of Diabetic candidates from huge Data using R-Studio	Sharmila K, Dr. S. A. V. Manickam	International Journal of Advanced Engineering Research and Science (IJAERS).	Decision tree	i. Using R, the dataset is analyzed and the correlation coefficient for two attributes is calculated.; ii. Predict the type of Diabetes by using decision tree based classification.	Possibility of developing efficient predictive models using the information from the analysis which is already carried out.

5	Identification of diabetes by using Classification techniques	A. Iyer, Jeyalatha S, and R. Sumbaly	International Journal of Data Mining & Knowledge Management Process (IJDKP).	i. Decision tree; ii. Naïve Bayes.	J48 Cross validation-74.8698%; J48 Percentage Split-76.9565%; Naive Bayes-79.5652%	i. In future, authors gather the data from distinct places all over the universe.; ii. The existing work can be enhanced for the diabetes analysis automatically.
6	A practical study on Sickness decision using Data extraction methods.	Deepika M, Dr. K. Kalaiselvi	International Conference on Inventive Communication & Computational Technologies (ICICCT).	i. Artificial Neural Network; ii. Decision Tree; iii. Logistic Regression; iv. Naïve Bayes; v.SVM	ANN: 73.23%; Logistic Regression: 76.13%; Tree based classifier:77.87%	Efficient and Accurate classifier can be developed.

Proposed System

The propose a classification model with boosted accuracy to predict the diabetic patient. In this model, we have employed different machine learning techniques are using like classification, regression and clustering. The major focus is to increase the accuracy by using resample technique on a benchmark well renowned PIMA diabetes dataset that was acquired from UCI machine learning repository, having eight attributes and one class label.[4] The proposed framework is shown in Figure 1.

The description of each phase is mentioned below:

Data Selection

Data selection is a process in which the most relevant data is selected from a specific domain to derive values that are informative and facilitate learning. PIMA diabetes dataset having 8 attributes that are used to predict the diabetes at earlier stage. This dataset is obtained from UCI repository.

Data Pre-processing

It also includes machine learning technique that includes changing crude information into reasonable configuration. The various activities involved here are Cleaning the data, integrating the data and transforming the data and Discretizing the data.[5]

Feature Extraction through Principle Component Analysis

Feature Extraction on the dataset to determine the most suitable set of attributes that can help achieve better classification. The set of attributes suggested by the PCA are termed as feature vector. Feature reduction or dimensionality reduction will be benefitted us by reducing the computation and space complexity.

Resampling Filter

The supervised Resample filter is applied to the pre-processed dataset. Resampling is a sequence of techniques used to rebuild your sample data sets, combining training sets and validation sets. In this study, Boot strapping resampling technique to enhance the accuracy.[7]

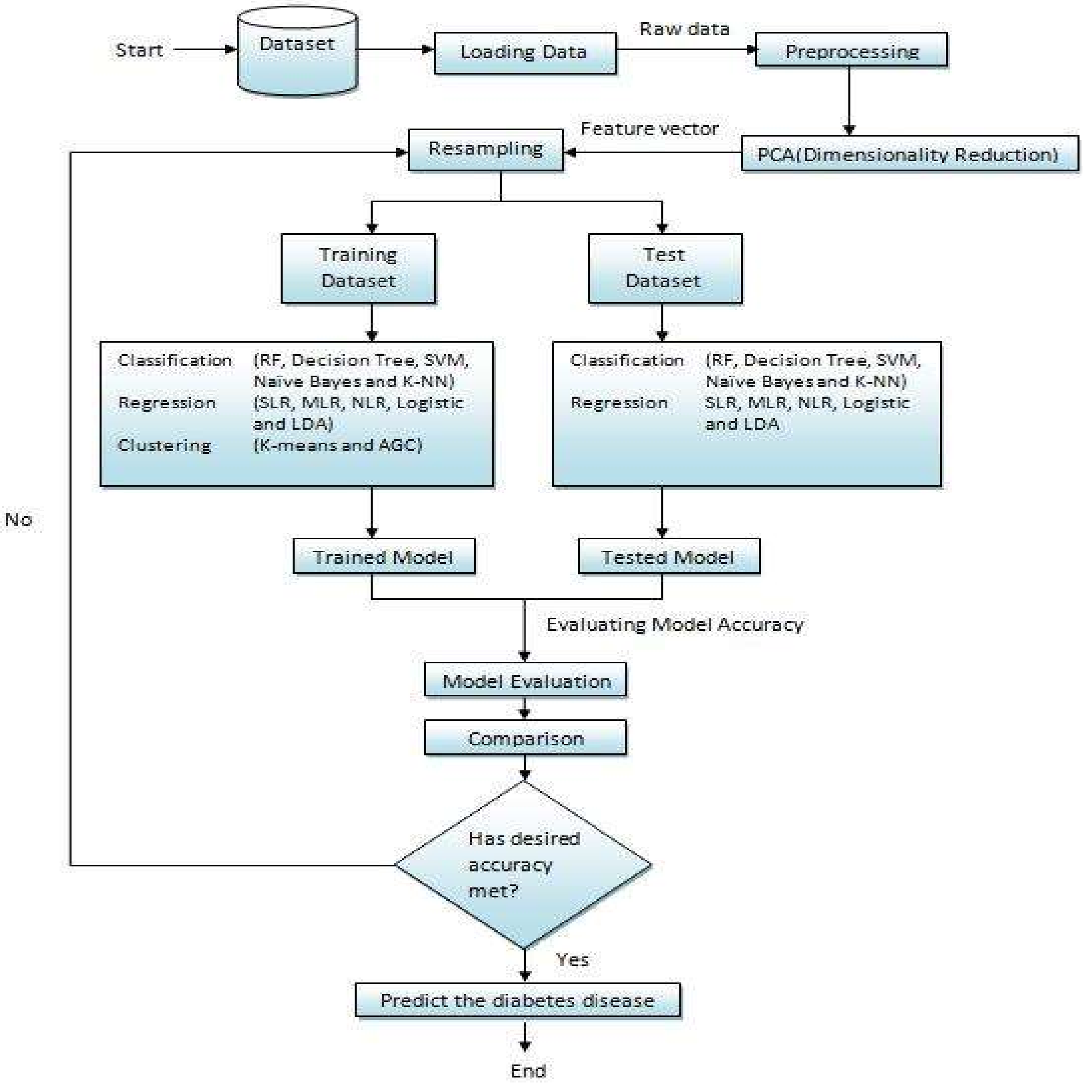
Start
Dataset
Loading Data
Raw data
Preprocessing
PCA(Dimensionality Reduction)
Feature vector
Resampling
Training Dataset
Test Dataset
Classification (RF, Decision Tree, SVM, Naïve Bayes and K-NN)
Regression (SLR, MLR, NLR, Logistic and LDA)
Clustering (K-means and AGC)
Classification (RF, Decision Tree, SVM, Naïve Bayes and K-NN)
Regression SLR, MLR, NLR, Logistic and LDA
No
Trained Model
Tested Model
Evaluating Model Accuracy
Model Evaluation
Comparison
Has desired accuracy met?
Yes
Predict the diabetes disease
End

Figure 1: Proposed System for Diabetes Prediction System

Machine Learning Techniques

Classification

Random Forest Technique

Random Judgement Forest is an outfit knowledge technique used for supervised learning methods and extra jobs that operates by building a group of decision trees during learning phase and label the class that is the frequency of the class labels used in classification and mean can be used for prediction of the independent trees. Irregular choice woods right for choice trees tendency for overfitting to their given known dataset.[8]

Support Vector Machines (SVM)

It is one of the most partition of classification Algorithms family. It is utilized strategy to implement prediction, guessing a class label and outlier analysis of data. SVM will grouping the information dependent on the hyper plane. The hyper plane should totally isolate the two class labels in the better way and the most extreme edge hyper plane ought to be picked as a best divider. The various classes of SVM are as follows:

i. Linearly separable Classifier
ii. Non-Linearly separable Classifier

Decision Tree

Decision tree algorithm is mainly used to produce a classification on training data and regression model into a tree structure, which is based on previous data to classify/predict class or target variables of future/new data with the help of decision rules or decision trees.[10] It can be useful for both numerical and categorical data. In a complete decision tree, the root node in each level is a starting point or the best splitting attribute in that position which helps to test on an attribute. The yield of the test will create branches. Leaf hub will go about as a last class mark or target variable to characterize/foresee the new information. Arrangement rules are attracted from root to leaf.

Naïve Bayes Method

The naive Bayes algorithm performs classification tasks in the field of ML. It can perform classification very well on the dataset even it has huge records with multi class and binary class classification problems.[11] The main application of Naive Bayes is text analysis and Natural Language Processing. Bayes theorem works based on conditional probability.

It can be represented as: P (X | Y) = [P (Y | X) * P (X) / P (Y)]

Here X and Y are two events and, P (X|Y) is the conditional probability of X given Y. P(X) is the probability of X.

P(Y) is the probability of Y. P (Y|X) is the conditional probability of Y given X.

k-Nearest Neighbors

K-NN is also a supervised learning model used for classification. It is also a significant choice for the classification type of analyses. In order to detect the target class label of a supplied data, this method detects variation between nearest given data classes and new given data values with the concern of K-value. [12] It uses K feature value in between 0-1usually.

Regression

Simple Linear Prediction Model

It explains the association between individual and dependent variables to detect the outcome of the dependent feature. A Simple prediction model uses only one individual feature.[13]

The simple linear regression model is represented as

y= (b0 +b1x)

Here, x (independent variable) and y (dependent variable) are two factors involved in simple linear regression analysis. Also, b0 is the Y-intercept and b1 is the Slope.

Multiple Linear Model Regression

It explains the association between two or more independent features and a dependent feature to guess the labels of a dependent feature.[14] It will take two or more individual variables. Dependent feature has a continuous but an independent variable may be discrete or continuous values.

The multiple linear regression model is represented as

y= (p0 +p1x1+p2x2+…+pnxn)

Here x1, x2,..., xn (independent variable) and y (dependent variable) are two factors involved in multiple linear regression analysis. Also, b0 is the y-intercept and p1, p2,…, pn is the slope.

Logistic based Prediction

It is one type of binary classification method and is used when the dependent feature is nominal. It models the association between one dependent feature and one or more independent variables.[15]

The various types of Logistic Regression are:

a. Binary class labeled Regression
b. Multinomial Regression
c. Ordinal type Regression

The categorical response has only two possible outcomes. Multinomial Logistic Regression has three or more outcomes without ordering whereas Ordinal Logistic Regression has three or more outcomes with ordering.

Polynomial Regression

Polynomial Regression is a type of prediction analysis which explains the association between the individual variable and dependent feature as polynomial of degree of ‘n’. It models a non-linear association between the independent variable and restrictive average of dependent variable. It should be represented as

x = a + b * y ^ n

Here ‘x’ is Dependent Variable, ‘y’ is Independent Variable and ‘n’ is Degree.

It is used to fit the data very well when the data is below and above the regression model. It minimizes the cost function and provides optimum result on the regression.[16]

Linear Discriminant Analysis (LDA)

It is a process of using different data elements and reflecting different functions to detect class labels of objects or instances separately. [17] It is applicable in the fields of Pattern Recognition and Prediction analysis.

Clustering

Mean based Clustering for the given ‘K’ Clusters

K-means grouping is not a supervised system learning algorithm. It can be used to solve clustering problems by classifying the dataset into a number of clusters k (group of similar objects), which defines the number of clusters which is assumed before classifying the dataset.[18]

Hierarchical Clustering

It is a type of clustering which is used to build a ladder of groups. The two methods of this type of grouping techniques are Agglomerative nesting and Divisive analysis. The first method is used to group objects into clusters based on their similarity. The result obtained at last is a tree representation of objects called 'Dendrogram'. Divisive Analysis is a best down methodology where all perceptions begin in one bunch, and parts are performed recursively as one moves down the pecking order. A hierarchical clustering is often represented as a dendrogram. Each cluster will be representing with centroids. Distance will be calculated by using linkage.[19]

Results and Discussions

Indian diabetes dataset named PIMA were used for analysis for this study. It consists of eight independent attributes and one independent class attribute. The study was implemented by R programming language using R Studio. Machine learning algorithms like classification (Judgement Tree, Bayesian Classification, KNN and Ensemble Learning), regression (linear, multiple, logistic, LDA) and clustering (k-means, hierarchical agglomerative) are used to predict the diabetics disease in early stages. Measure Performance model by using accuracy.

Table 1: Predictive Analysis of Machine Learning Algorithms

S. No	Algorithm	Accuracy
1	Random forest	73%
2	Decision tree	67%
3	SVM	82%
4	Naïve Bayes	76%
5	K-NN	81%
6	Simple linear regression	88%
7	Logistic regression	78%
8	LDA	78%
9	k-Means	71%
10	Hierarchical agglomerative	64%

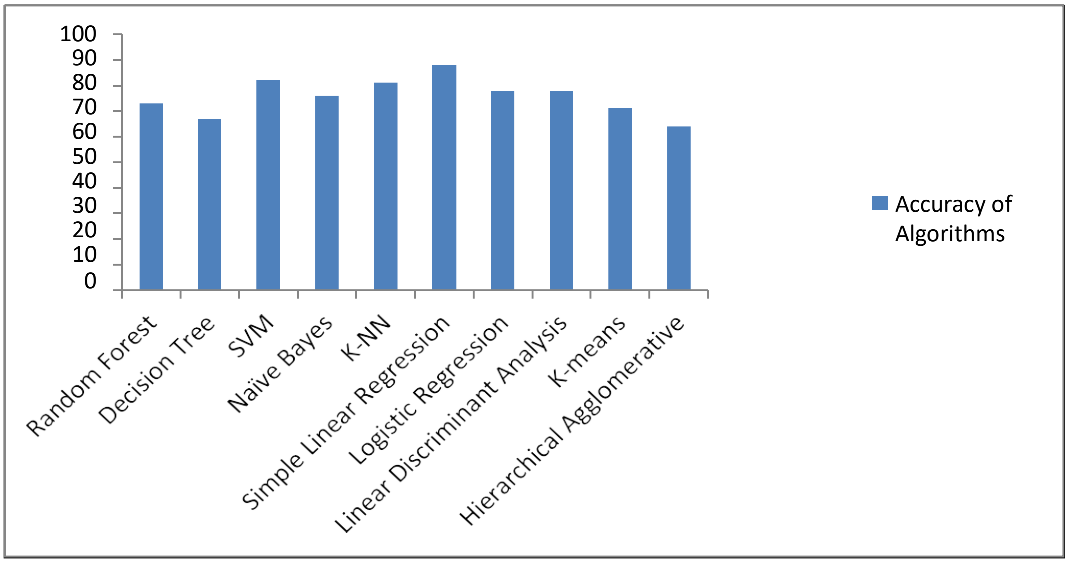

Figure 2: Comparison of Accuracy of Various Algorithms

Conclusion and Future Work

Knowledge extraction creates a significant position in various fields such as Simulated Intelligence and Device Learning, Database Systems and more. The core objective is to enhance the accuracy of predictive model**.** This PIMA dataset will increase the accuracy of almost all algorithms but the SVM and linear regression leads over others. In future many advanced techniques will be used to increasing the accuracy of the algorithms.

References

[1] Dr. O. Obulesu, M. Mahendra and M. Thrilok Reddy, “Machine Learning Techniques and Tools: A Survey”, Proceedings of the IEEE International Conference on Inventive Research in Computing Applications (ICIRCA 2018) at RVS College of Engineering & Technology, Coimbatore, Tamilnadu during 11-12 July, 2018, IEEE Xplore Compliant Part Number:CFP18N67-ART; ISBN:978-1-5386-2456-2, Page No’s: 618-624, ISBN: 978-1-5386-2456-2/18©2018 IEEE.

[2] Jafar Tanha, “Semi-supervised self-training for decision tree classifiers”, International Journal of Machine Learning and Cybernetics,Volume 8, Issue 1, Page No’s: 355-370, January, 2015.

[3] Khadim D, Fleur M and Gayo D, “Large scale biomedical texts classification: a k-NN and an ESA-based approaches”, Journal of Biomedical Semantics, 7:40, June, 2016.

[4] Hong R, H. M. Wang and Jian L, “Privacy-Preserving k-Nearest Neighbor Computation in Multiple Cloud Environments, IEEE Access, ISSN: 2169-3536, Volume 4, Page No’s: 9589-9603, December, 2016.

[5] L. Jiang, C. Li, “Deep feature weighting for naive Bayes and its application to text classification”, Journal of Engineering Applications of Artificial Intelligence, Volume 52, Page No’s: 26-39, June, 2016.

[6] Ahmed M, Alison H, “Modeling built-up expansion and densification with multinomial logistic regression, cellular automata and genetic algorithm”, Volume 67, Page No’s: 147-156, January, 2018.

[7] T. Razzaghi, Oleg R, “Multilevel Weighted Support Vector Machine for Classification on Healthcare Data with Missing Values”, PLUS ONE, Page No’s:1-18, May 2016.

[8] Hui L, D. Pi, “Integrative Method Based on Linear Regression for the Prediction of Zinc binding Sites in Proteins”, IEEE Access Volume 5, Page No’s: 14647-14656, August, 2017.

[9] L. Wang, D. Wang, “Intelligent CFAR Detector Based on Support Vector Machine”, IEEE Access, Volume 5, Page No’s: 26965-26972, December, 2017.

[10] Enrico R, Michel L, “The Counter, a Frequency Counter Based on the Linear Regression”, IEEE Transactions on Ultrasonics, Ferroelectrics Volume 63, Issue 7, Page No’s: 961-969, July, 2016.

[11] J. Han, M. Kamber and J. Pei, “Data Mining: Concepts and Techniques”, 3rd Edition, MK Series, 2012.

[12] O. Obulesu and Dr. A. Rama Mohan Reddy, “Finding Frequent and Maximal Periodic Patterns in Spatio-temporal Databases for Shifte Instances”, CiiT Data Mining and Knowledge Discovery, Volume 6, Issue no.5, Page No’s: 224-232, June, 2014.

[13] O. Obulesu and Dr. A. Rama Mohan Reddy, “Finding Maximal Periodic Patterns and Pruning Strategy in Spatiotemporal Databases” International Journal of Advanced Research in Computer Science and Software Engineering, Vol. 2, Issue No. 4, Page No.s:423- 426, April 2012.

[14] Xindong Wu, Xingquan Zhu, Gong-Qing Wu and Wei Ding “Data Mining with Big Data”, IEEE Transactions on Knowledge and Dat Engineering, Vol. No. 26, Issue No.1, January, 2014.

[15] Dr. O. Obulesu and Dr. A. Rama Mohan Reddy, “Fast and Efficient Mining of Frequent and Maximal Periodic Patterns in Spatiotempora Databases for shifted instances”, International Conference on Advanced Computing (IACC-2016) at SRKR Engineering College Bhimavaram, A.P., ISBN: 978-1-4673-8286-1/16 $31.00, DOI: 10.1109/IACC.2016.17, Page No’s: 35-40.

[16] O. Obulesu, Dr. A. Rama Mohan Reddy and K. Suresh, “Finding Maximal Periodic Patterns and Pruning Strategy in Spatiotempora Databases”, International Journal of Advanced Research in Computer Science and Software Engineering, Vol.2, Issue No. 4, April, 2012 Page No’s: 423-426, ISSN: 2277-128X.

[17] Mr. O. Obulesu and Ms. O. Gireesha, “TKAR: Efficient Mining of Top – k Association Rules on Real – life Datasets”, 5th International Conference on Frontiers of Intelligent Computing: Theory and applications organized by KIIT University, Bhubaneswar, Odisha during September 16-17, 2016.

[18] Mr.O.Obulesu, Dr.A.Rama Mohan Reddy and Mahendra M, “A Comparative study of Frequent and Maximal Periodic Pattern Mining Algorithms in Spatiotemporal Databases”, International Conference on Advanced Material Technologies (ICAMT-2016) at Dadi Institute Engineering & Technology, Anakapalli, Visakhapatnam during 27-28 December, 2016.

[19] S. Abdul, Suresh Babu D and O. Obulesu, “Application of Big Data Analytics in Power System under Single Transmission Line Outage Condition”, Proceedings of the Third International Conference on Trends in Electronics and Informatics (ICOEI 2019) at SCAD College of Engineering & Technology, Tirunelveli, Tamilnadu during 23-25 April, 2019, IEEE Xplore Compliant Part Number: CFP19J32-DVD, ISBN:978-1-5386-9438-1, Page No’s: 520-525.