ISSN 2347 - 3983
Volume 8. No. 7, July 2020
International Journal of Emerging Trends in Engineering Research
Available Online at http://www.warse.org/IJETER/static/pdf/file/ijeter137872020.pdf
https://doi.org/10.30534/ijeter/2020/137872020

Disease Prediction and Performance Analysis using Data Analytics

Shruthi J[1], Mamatha Bai B G[2]
[1]M. Tech Scholar, Computer Science & Engineering, Nitte Meenakshi Institute of Technology, Bangalore, India, shruthij1997@gmail.com
[2]Assistant Professor, Computer Science & Engineering, Nitte Meenakshi Institute of Technology, Bangalore, India, mamathamane@gmail.com

ABSTRACT

Big Data means the large and the voluminous data collected from various resources. The data helps us in understanding lot many aspects of the particular field. Healthcare is one such field or area which produces huge number of data such as patient's profiles, disease, symptoms and so on. Data Analytics comes into the picture when we have to review, analyze and predict the data. The main objective of the Data Mining Techniques is to make smart decisions, consume less time and being more accurate. In this paper, we have used two different datasets: Pregnancy Diabetes Dataset and Breast Cancer Dataset. The Classification Algorithms like SVM and Random Forest for the top-level classification and Clustering Algorithms like Agglomerative, Gaussian Mixture Model, and Spectral Clustering for further processing. The efficient algorithm is identified by comparing the accuracy results of the different implemented clustering algorithms.

Key words: Agglomerative, Big Data Analytics, Breast Cancer, Data Mining, GMM, Healthcare, Pregnancy Diabetes, Random Forest, Spectral , SVM,.

1. INTRODUCTION

Big Data is one of the main areas that needs more focus and more research activities are taking place. Big data is the combination and collection of structured, unstructured and semi-structured data [1]. In the current world full of technology and innovations, Big Data Analytics have achieved the at most importance and it is the developing technology leading for creating a range of curiosity with the advancements in the application areas. With innovation progression, the period of information in different technicalities, is developing liberally and taking care of those information may be a major summons in genuine constraint. It bargains with the exclusive degree of information and examination, which retrieves the result instantaneously. As we all know, it is a vast collection of information & bewildered information that can be intense to bargain with ordinary actualities. Big Data rearranges the unorganized information to the productive experiences that is advantageous in keeping up the mind blowing sum of information received through different sources. This thought of managing the expansive sum of information is not modern but there could be a better approach in handling it. Healthcare in Big Data is one of the most important fields which have worked with the information that deals with the 5 V's that's Velocity, Variety, Volume, Veracity and Value which are natural components of the realities. The truth is unfurling among a few of healthcare, composition, wellness, insurers, researcher's rights, organization etc. The characteristic complexity of the healthcare truths the advantage in creating and implementing big data caches and detects the arrangements to the issues that are being raised. Traditional forms of science have usually targeted the examination of diseases which are based on the modifications within the form of confined insights of a particular methodology of information.

1.1 Importance of Big Data in Healthcare

The role of Big Data in Healthcare is gaining importance corresponding to the volume of information which is either structured or unstructured. Medical data is confidential and ought to be taken care of. Analysis of the diseases based on the symptoms makes a difference where the specialists and patients induce to know about the wellbeing conditions at an early stage and accordingly follow preventive measures. Big data provides the strategy of standard information frameworks [2]. The data is too much colossal and dealing with the genuine truths can be a great challenge within the field of Healthcare. Dealing with huge data implies to electronic wellbeing and is troublesome to give away with ancient or essential data administration methods and old-fashioned bundle.

1.2 Pregnancy Diabetes

Gestational diabetes is diabetes analyzed during preliminary time of pregnancy. Like other types of diabetes, gestational diabetes influences how cells utilize glucose. Gestational diabetes causes high level blood sugar that can influence the pregnancy and the fetus.
Those who create gestational diabetes are at higher hazard of creating type-2 diabetes in future. In most cases, there are no indications. A blood sugar test during pregnancy is

diagnosed. Treatment procedures incorporate day by day blood sugar observing, a solid count calories, work out and checking the infant. On the off chance that blood sugar is as well high, pharmaceutical assistance is required.

The diabetes during pregnancy appears in 3 different stages: “Low”, “Mid” and “High”. For a few ladies, diabetes start and end during their pregnancy period. But for few, it is carried throughout their life. Some of the babies do get affected from the mother’s womb itself.

1.3 Breast Cancer

Breast cancer gets created in breast cells. Regularly, the cancer shapes in either the lobules or the channels of the breast. Lobules are the organs that create drain, and channels are the pathways that bring the drain from the organs to the areola. Breast cancer can happen in ladies and seldom in men. Symptoms of breast cancer incorporate a protuberance within the breast, ridiculous release from the nipple and changes within the shape or surface of the areola or breast. Its treatment depends on the stage of cancer. It may comprise of chemotherapy, radiation, hormone treatment and surgery. The anatomy of the breast cancer tumor can be seen in the Figure 1.

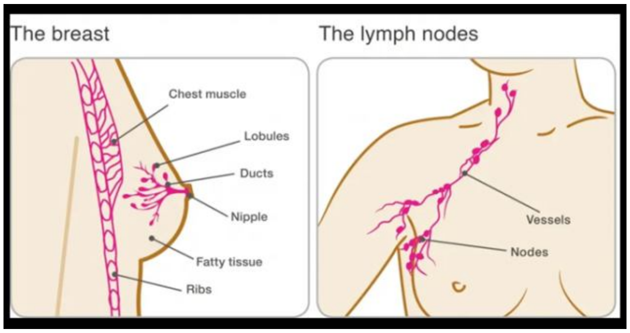


Figure 1: The anatomy of Breast Cancer and the Tumor

Breast Cancer occurs in 4 different stages, named such as “Pre-cancer”, “Intermediate Cancer”, “Advanced Cancer” and “Meta-static Cancer”, which is shown in the Figure 2.

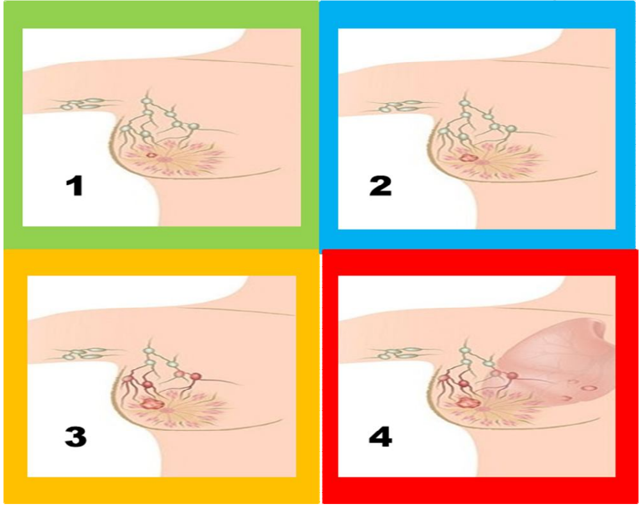


Figure 2: Four stages of Breast cancer

2. LITERATURE SURVEY

In [3] Data mining approaches offer the methodology and technology to transform these heterogeneous data into meaningful information for decision making. Large number of healthcare dataset is considered. Keeping k means as reference, G means algorithm was developed. Clustering methodologies are very efficient. Both algorithm and data taken should be pre-processed using specific rules so that they can enhance the performance and the results of the clustering algorithm. [4] GMM is the simplest algorithm which gives the efficient result and accuracy. GMM uses good number of parameters to get the result out of the trained data. [5] Random Forest has given the better accuracy. It has the advantage of computing the importance of each variable in the classification process. In the same way for our work we have used healthcare dataset, adopted Random Forest and SVM for the classification. AdaBoost, Random Forest was used as base classifier (93.3%) [6].

In [7] Breast cancer is studied and spectral clustering is implemented. It is adopted in our work for detecting he stages of the disease. Spectral clustering can be achieved directly. [8] explains Constrained Spectral Clustering with the performance of a series of classic models in terms of classification accuracy with respect to whether individual’s reported symptoms increase or decrease over time Health care has been taken as the work. In [9], the average accuracy of proposed model is 85.59% which is higher as compared to other models proposed earlier. . SVM classification algorithm is used to classify the data into different number of classes. The algorithms like EM, ANN, C4.5 have contributed in the analysis of data on kidney disease which referred as paper [10]. The algorithms have proven themselves on contribution by classifying the data into accurate classes and also providing the great accuracy of 70%, 75% and 96.5% respectively. Mamatha Bai B G et. al. [11] has focused on Gaussian Naïve Bayes, OPTICS and BIRCH algorithms for the analysis of Diabetes. Analysis of these algorithms is done by comparing the accuracy and by changing the various metrics of the performance. And concludes the OPTICS Algorithm gives the best result and is suitable for the given dataset.

The publication on the Data Mining Technology for Video Summarization can be seen in the paper [12] which makes use of CURE algorithm for both big data and video summarization streams. The algorithm gives the best results for both the streams. The paper [13] is aimed to study and compare the effect of bagging on classification accuracy by using different decision trees as the base classifiers. The experiment shows the effect of bagging on various base classifiers. The various statistical studies carried out for different algorithms for medical data is presented in [14].

Madam Chakradar et. al. [15] has explained about Support Vector Machine, Logistic Regression and other algorithms for Diabetes Mellitus of Type 2. The paper explains about the finding insulin through Machine Learning Algorithms. The accuracy of 97% is achieved. In [16], the author explains about the mobile application which will help the normal people to keep track on their blood pressure, using microcontrollers and also using sensors to detect the blood pressure.

3. PROPOSED METHODOLOGY

This work predicts the presence of Diabetes and Breast Cancer and performance analysis of various algorithms implemented. Figure 3 represents the work flow of the proposed method. The figure describes the whole architecture of the system, the flow of the algorithms from the data pre-processing to the results. The proposed methodology uses both classification and clustering technologies.

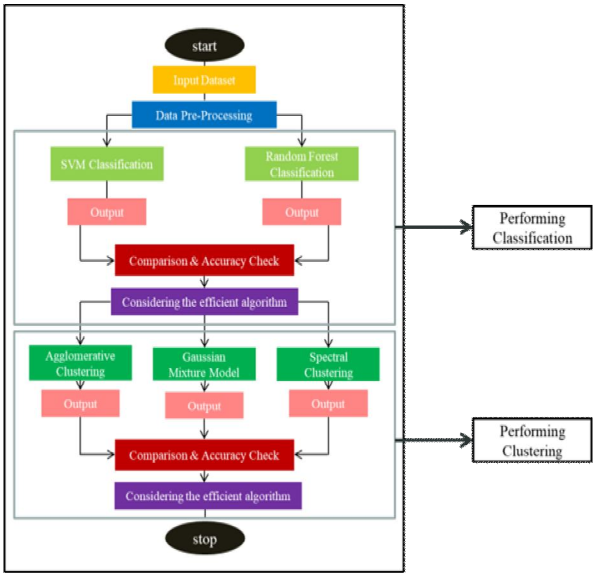

start
Input Dataset
Data Pre-Processing
SVM Classification
Random Forest Classification
Output
Output
Comparison & Accuracy Check
Considering the efficient algorithm
Performing Classification
Agglomerative Clustering
Gaussian Mixture Model
Spectral Clustering
Output
Output
Output
Comparison & Accuracy Check
Considering the efficient algorithm
Performing Clustering
stop

Figure 3: Proposed Methodology

The dataset is read in the first step and pre-processed for the missing values. Once it is done, the classification techniques are applied which further forms the classes based on the dataset for top level classification. Further the affected dataset is fed to the clustering algorithms for the deeper level clustering, where the levels and the stages of the disease is being predicted and further performance analysis is done.

3.1 Dataset Description

A. Pregnancy Diabetes Dataset

The dataset is taken from the source [17]. The dataset contains 1000 records of the patients with 9 varied attributes, which is listed in the Table 1.

Table 1: Attributes of Pregnancy Diabetes Dataset

Sl. No.	Attributes
1.	Patient Id
2.	Plasma Glucose Level
3.	Diastolic Blood Pressure
4.	Triceps Thickness
5.	Serum Insulin
6.	Body Mass Index
7.	Diabetes Pedigree
8.	Age
9.	Class

B. Breast Cancer Dataset

The dataset is taken from the source [18]. The dataset contains 10,000 records of the patients with 10 varied attributes, which are calculated for the values listed in the Table 2.

Table 2: Attributes of Breast Cancer Dataset

Sl. No.	Attributes
1.	Radius
2.	Texture
3.	Perimeter
4.	Area
5.	Smoothness
6.	Compactness
7.	Concavity
8.	Concave points
9.	Symmetry
10.	Fractal Dimensions

3.2 Algorithms

A. Support Vector Machine (SVM)

Algorithm: SVM
Input: Dataset to classify whether Diseased or healthy
Output: Two Classes predicting the presence of Disease
Steps:

1. At first, what SVM does is to find the line which separates the two classes.
2. The taken dataset is fed to the algorithm
3. Finds out the data points that are the closest to the line from both the classes. Maximizing the margin is the major goal, which is called as the hyper plane.
4. $\mathbf{z = x^2+y^2}$, **z** coordinate is the squared distance of the data point from the origin. We add the z coordinate to make the data points linearly separable.
5. Thus we can keep adding the extra dimensionalities to make it linearly separable.

B. Random Forest Classification

Algorithm: Random Forest
Input: Dataset to classify whether Diseased or healthy
Output: Two Classes predicting the presence of Disease
Steps:

1. Input the given dataset
2. Divide the dataset into training and testing data
3. Selection of samples randomly for the given datasets.
4. Decision tree is constructed for every sample by the algorithm; prediction results from each decision tree are retrieved.
5. Algorithm will perform voting for every predicted result.
6. The most voted result is decided as the best solution for the dataset given.

C. Agglomerative Clustering

Algorithm: Agglomerative Clustering
Input: Dataset to predict the stages of the disease
Output: Three clusters predicting the stages of Disease
Steps:

1. The data set given will have (p1, p2, p3, …………, pn) of size N
2. It computes the distance matrix
3. for (i=1 to N)
4. for (j=1 to N)
5. distance_matrix [i] [j] = dis [p1, p2] here, each point is considered as individual clusters.
6. Now, repeat the steps for two clusters to merge into one, which has the closest distance.
7. **7.** Repeat until all the clusters are being merged.

D. Gaussian Mixture Model (GMM)

Algorithm: GMM
Input: Dataset to predict the stages of the disease
Output: Three clusters predicting the stages of Disease
Steps: The Steps of the Gaussian Mixture Model can be seen from the Figure 4.

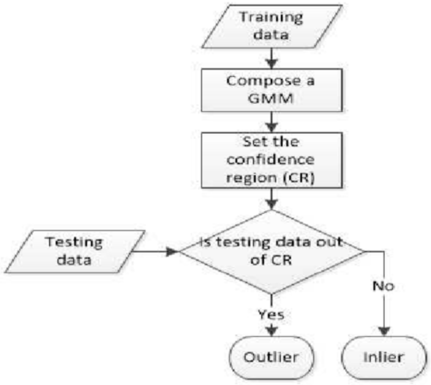


Figure 4: Representing the workflow of the GMM

E. Spectral Clustering

Algorithm: Spectral Clustering
Input: Dataset to predict the stages of the disease
Output: Three clusters predicting the stages of Disease
Steps:

1. Building similarity graph using:
   a. Epsilon
   b. K-Nearest Neighbors
   c. Fully Connected
2. Predict the data onto a lower Dimensional Space
3. Clustering of the Data

4. EXPERIMENTAL RESULTS AND ANALYSIS

The dataset was fed to the classification algorithms, where the first level of classification is done by dividing the dataset to two classes. For the Diabetic dataset, the classes are: Diabetic and Non Diabetic. For the Breast cancer, we have Malignant and Benign as two different classes. The corresponding graphs are represented in this section.

4.1 Pregnancy Diabetes Data Analysis

Classification for top level categories for Pregnancy Diabetes Dataset are done using SVM and Random Forest which can be seen in Figures 5 and 6.

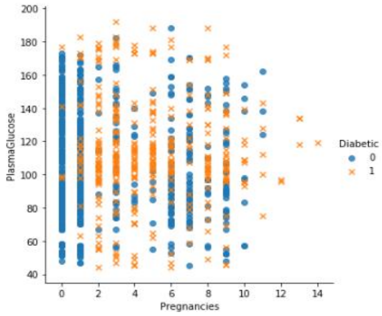


Figure 5: O/p of SVM classifier for Diabetic Dataset

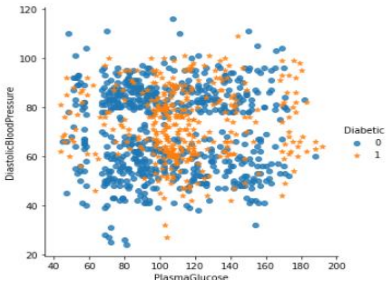


Figure 6: O/p of Random Forest classifier for Diabetic Dataset

Further, we proceed to find out the severity stages by using the clustering algorithms, where the clusters are formed using the clustering algorithms.
Clusters were formed based on the stages of Diabetes such as

- **Low**
- **Mid**
- **High**

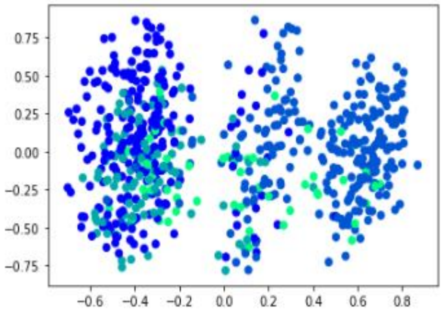


Figure 7: O/p of Agglomerative Clustering for Diabetic Data

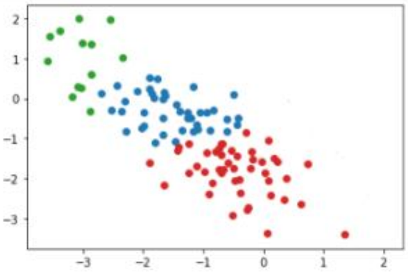


Figure 8: O/p of GMM for Diabetic Data

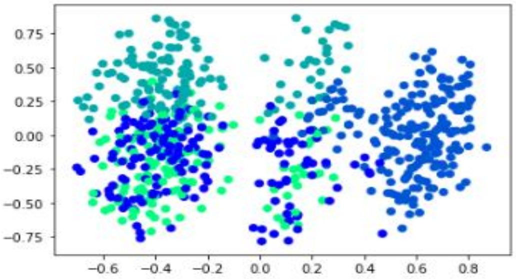


Figure 9: O/p of Spectral Clustering for Diabetic Data

Each color in the graphs denotes different clusters based on the stages. We can notice in Figures 7, 8 and 9 that there are three different clusters denoting the three different stages of the diabetes.

4.2. Breast Cancer Data Analysis

Classification for top level categories for Breast Cancer Dataset are done using SVM and Random Forest which can be seen in Figures 10 and 11.

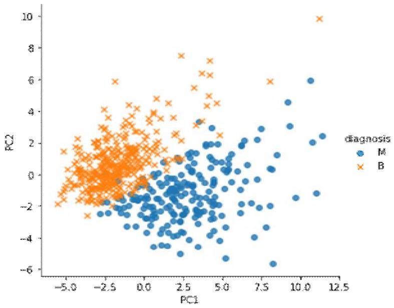


Figure 10: O/p of SVM classifier for Breast Cancer Data

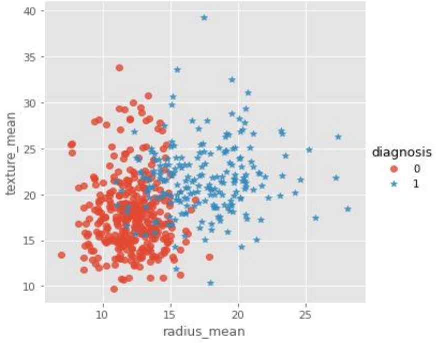


Figure 11: O/p of Random Forest classifier for Breast Cancer Data

Further, we proceed to find out the severity stages by using the clustering algorithms, where the clusters are formed using the clustering algorithms.

Clusters were formed based on the stages of breast cancer such as:

- **Stage-1: Pre-Cancer**
- **Stage-2: Intermediate Cancer**
- **Stage-3: Advanced Cancer**
- **Stage-4: Meta-static Cancer**

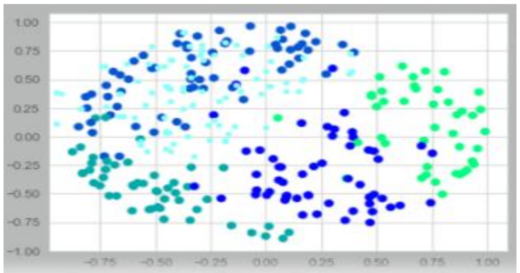


Figure 12: O/p of Agglomerative for Breast Cancer Data

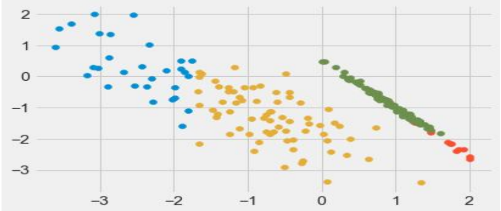


Figure 13: O/p of GMM for Breast Cancer Data

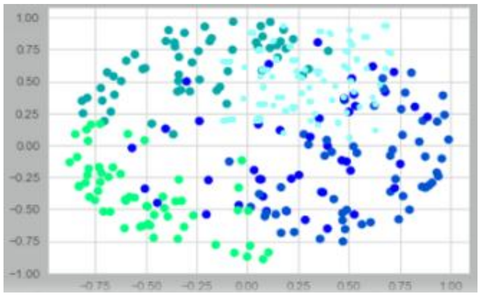


Figure 14: O/p of Spectral clustering for Breast Cancer Data

Each color in the graphs denotes different clusters based on the stages. We can notice in Figures 12, 13 & 14 that there are three different clusters denoting the three different stages of the diabetes.

5. STATISTICAL ANALYSIS

Table 3 represents the accuracy of all the algorithms that are placed against each other for the better understanding.

We can observe that from the classification algorithms, SVM gives the better results for both the datasets, and in the clustering algorithms GMM method gives us the better accuracy.

Table 3: Accuracy of Algorithms

Algorithm	Diabetes	Breast Cancer
SVM	96.4%	97.2%
Random Forest	96.1%	96.5%
GMM	81%	79%
Agglomerative	67%	65.6%
Spectral	61%	62.3%

5.1. Statistical analysis of the pregnancy diabetes dataset.

Figure 15 & 16 represent the classification algorithms results. The percentage representations of the division of categories are depicted using the graphical visualization.

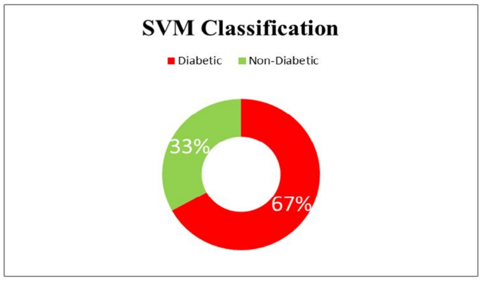


Figure 15: Graphical representation of the prediction results obtained by SVM for Diabetic Data

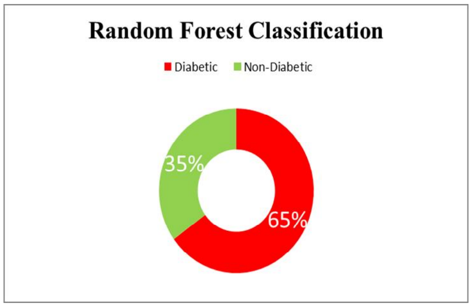


Figure 16: Graphical representation of the prediction results obtained by Random forest for Diabetic Data.

The prediction results obtained by different Clustering Algorithms for Diabetic Data is graphically represented in Figure 17.

5.2 Statistical analysis of the breast cancer dataset.

Figures 18 & 19 represent the classification algorithms results. The percentage representations of the division of categories are being depicted using the graphical visualization.

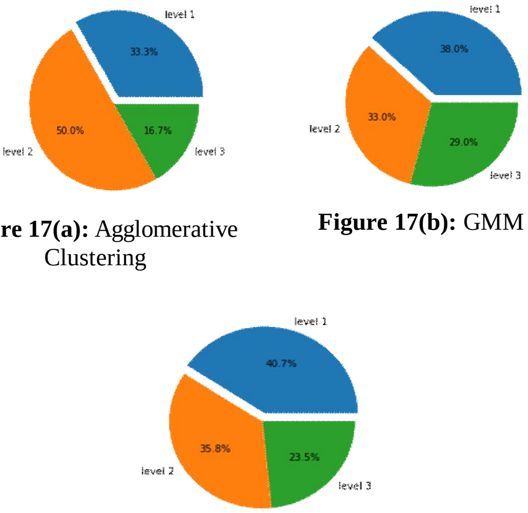


Figure 17(a): Agglomerative Clustering

Figure 17(b): GMM

Figure 17(c): Spectral Clustering

Figure 17: Graphical Representation of Prediction results obtained by different Clustering Algorithms for Diabetic Data

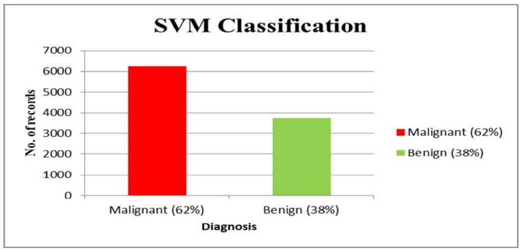


Figure 18: Graphical representation of the prediction results by SVM for Breast Cancer

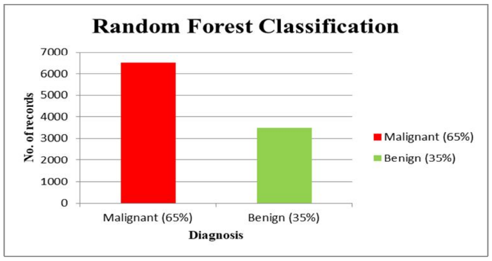


Figure 19: Graphical Representation of the prediction results by Random Forest for Breast Cancer

The statistical analysis for Breast Cancer using clustering algorithms is illustrated in Figures 20, 21 and 22.

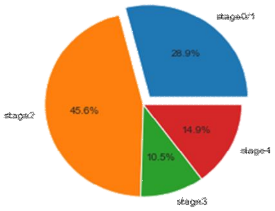


Figure 20: Graphical representation of the prediction results obtained by Agglomerative Clustering for Breast Cancer

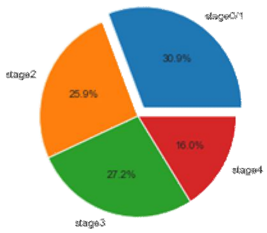


Fig 21: Graphical representation of the prediction results obtained by GMM for Breast Cancer

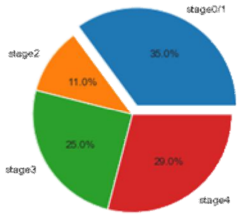
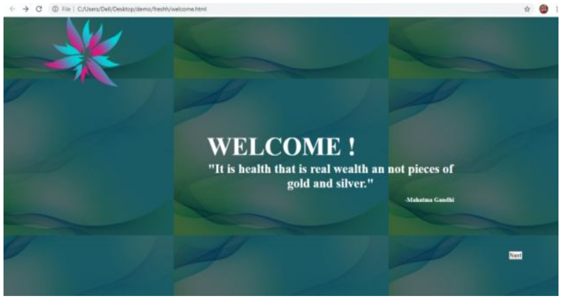


Figure 22: Graphical representation of the prediction results obtained by Spectral Clustering for Breast Cancer

6. WEB APPLICATION RESULTS

The user interface is the one that we can experience and use it as platform for sharing information. Having this thought, the results obtained from the algorithms, their accuracies, statistical results are constituted using HTML and CSS technology, where one can use the website to refer the results and analyze further.

Figure 23 shows the Welcome Page of the Website developed followed by the Information Page in Figure 24. Classification and Clustering Results in the UI environment are shown in Figures 25 and 26 respectively. Figure 27 and 28 depicts the Accuracy and Statistical analysis of our work.

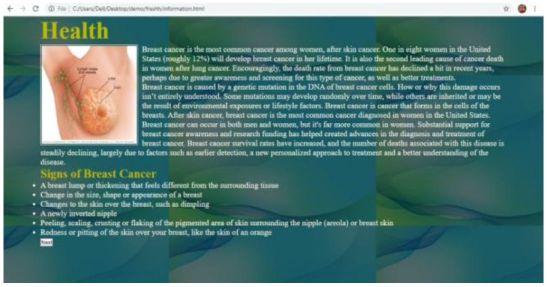


WELCOME !

"It is health that is real wealth an not pieces of gold and silver."

-Mahatma Gandhi

Figure 23: Welcome Page



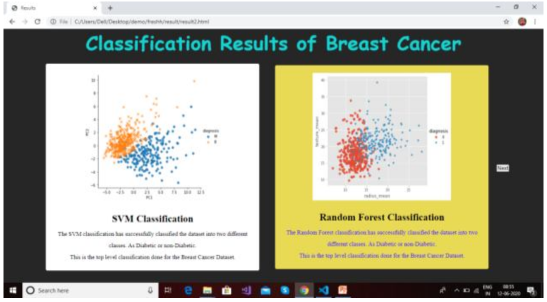
Health

Breast cancer is the most common cancer among women, after skin cancer. One in eight women in the United States (roughly 12%) will develop breast cancer in her lifetime. It is also the second leading cause of cancer death in women after lung cancer. Encouragingly, the death rate from breast cancer has declined a bit in recent years, perhaps due to greater awareness and screening for this type of cancer, as well as better treatments.
Breast cancer is caused by a genetic mutation in the DNA of breast cancer cells. How or why this damage occurs isn't entirely understood. Some mutations may develop randomly over time, while others are inherited or may be the result of environmental exposures or lifestyle factors. Breast cancer is cancer that forms in the cells of the breasts. After skin cancer, breast cancer is the most common cancer diagnosed in women in the United States. Breast cancer can occur in both men and women, but it's far more common in women. Substantial support for breast cancer awareness and research funding has helped created advances in the diagnosis and treatment of breast cancer. Breast cancer survival rates have increased, and the number of deaths associated with this disease is steadily declining, largely due to factors such as earlier detection, a new personalized approach to treatment and a better understanding of the disease.

Signs of Breast Cancer

- A breast lump or thickening that feels different from the surrounding tissue
- Change in the size, shape or appearance of a breast
- Changes to the skin over the breast, such as dimpling
- A newly inverted nipple
- Peeling, scaling, crusting or flaking of the pigmented area of skin surrounding the nipple (areola) or breast skin
- Redness or pitting of the skin over your breast, like the skin of an orange

Figure 24: Information Page



Classification Results of Breast Cancer

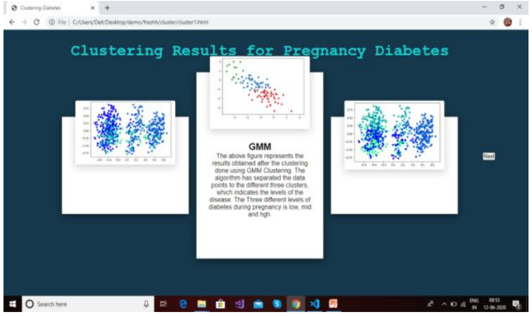
SVM Classification

The SVM classification has successfully classified the dataset into two different classes. As Diabetic or non-Diabetic.
This is the top level classification done for the Breast Cancer Dataset.

Random Forest Classification

The Random Forest classification has successfully classified the dataset into two different classes. As Diabetic or non-Diabetic.
This is the top level classification done for the Breast Cancer Dataset.

Figure 25: Classification Results in the UI

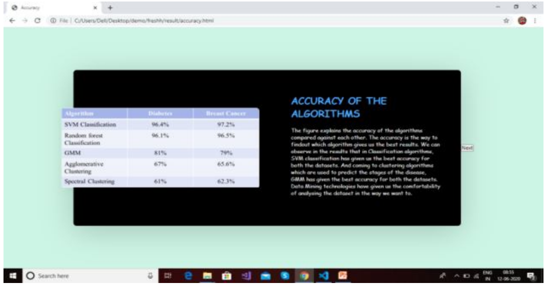


Clustering Results for Pregnancy Diabetes

GMM

The above figure represents the results obtained after the clustering done using GMM Clustering. The algorithm has separated the data points to the different three clusters, which indicates the levels of the disease. The Three different levels of diabetes during pregnancy is low, mid and hgh.

Figure 26: Clustering Results in the UI

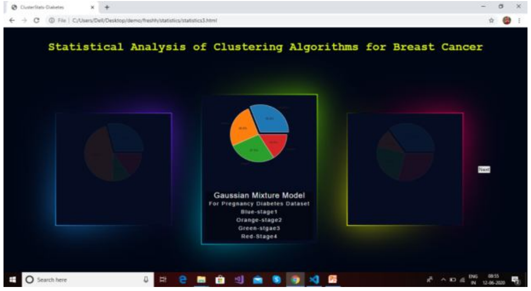


Algorithm	Diabetes	Breast Cancer
SVM Classification	96.4%	97.2%
Random forest Classification	96.1%	96.5%
GMM	81%	79%
Agglomerative Clustering	67%	65.6%
Spectral Clustering	61%	62.3%

ACCURACY OF THE ALGORITHMS

The figure explains the accuracy of the algorithms compared against each other. The accuracy is the way to findout which algorithm gives us the best results. We can observe in the results that in Classification algorithms, SVM classification has given us the best accuracy for both the datasets. And coming to clustering algorithms which are used to predict the stages of the disease, GMM has given the best accuracy for both the datasets. Data Mining technologies have given us the comfortability of analysing the dataset in the way we want to.

Figure 27: Accuracy page



Statistical Analysis of Clustering Algorithms for Breast Cancer

Gaussian Mixture Model
For Pregnancy Diabetes Dataset
Blue-stage1
Orange-stage2
Green-stgae3
Red-Stage4

Figure 28: Statistical Analysis page

7. CONCLUSION AND WAY FORWARD

The accuracy of the each algorithm is being compared against each other. We can observe that SVM classifier under classification algorithms for top level classification gives the highest accuracy for the datasets used. And the GMM model under clustering algorithms, gives the highest accuracy for both the datasets.

The main objective and motivation of this work is all about the women's healthcare, be it a diabetes or breast cancer and to create awareness among the Women Community especially in the rural sectors. Further, we can develop an android application which will be an Interactive platform for the users to know about the disease and also to create the public awareness.

ACKNOWLEDGEMENT

The authors express their sincere gratitude to Prof. N R Shetty, Advisor and Dr. H C Nagaraj, Principal, Nitte Meenakshi Institute of Technology for giving constant support to carry out research at NMIT.

REFERENCES

1. Scardapane, Rosa Altilio,Valentina Ciccarelli, Aurelio Uncini and Massimo Panella, "**Privacy-Preserving Data Mining for Distributed Medical Scenarios**", *00184 Rome, Italy of the Springer International Publishing AG* 2018. https://doi.org/10.1007/978-3-319-56904-8_12
2. B G Mamatha Bai, R Harshitha, Jharna Majumdar, "**Diagnosis of Cardiotocography for Analysis and Prediction of Treatment during Pregnancy**", *Proceedings of 2018 International Conference on Advances in Computing, Communications and Informatics*, IEEE Explore, ISBN: 978-1-5386-5314-2, DOI: 10.1109/ICACCI.2018.8554437.
3. Ramzi.Haraty, MohamadDimishkieh, and MehediMasud "**An Enhanced k-Means Clustering Algorithm for Pattern Discovery in Healthcare Data**" *Proceedings of the Industrial and Systems Engineering Research Conference*, 2018.
4. Felicity R Allen, Eliathamby Ambikairajah, Nigel H Lovell, and Branko G Celler, National Information and Communications Technology Australia, "**Classification of a known sequence of motions and postures from accelerometry data using adapted Gaussian mixture models**" Received 12 February 2016, accepted for publication 27 June 2016, Published 25 July 2017.
5. Khalilia, Sounak Chakraborty and Mihail Popescu, "**Predicting disease risks from highly imbalanced data using random forest**" *USAKhalilia et al. BMC Medical Informatics and Decision Making* 2018.
6. María N. Moreno García, Juan Carlos Ballesteros Herráez, Mercedes Sánchez Barba and Fernando Sánchez Hernández, "**Random Forest Based Ensemble Classifiers for Predicting Healthcare-Associated Infections in Intensive Care Units**" *13th International Conference, Advances in Intelligent Systems and Computing 474,* 2016
7. B. Walker, Jacob N. Norris, Anna E. Tshiffely, Melissa L. Mehalick, Craig A. Cunningham, Ian N. Davidson, Senior Member, "**Applications of Transductive Spectral Clustering Methods in a Military Medical Concussion Database**" *IEEE/ACM Transactions on Computational Biology and Bioinformatics Proceedings of the 2016 IEEE TRANSACTIONS ON COMPUTATIONAL BIOLOGY AND BIOINFORMATICS.* https://doi.org/10.1109/TCBB.2016.2591549
8. Chua Kuang Chua, Vinod Chandran, Rajendra U. Acharya and Lim Choo Min 1Department of Electronics and Computer Engineering "**Cardiac Health Diagnosis Using Higher Order Spectra and Support Vector Machine**", *The Open Medical Informatics Journal*, 2016, 3, 1-8.
9. Divya Tomar and Sonali Agarwal Indian Institute of Information and Technology, "**Feature Selection based Least Square Twin Support Vector Machine for Diagnosis of Heart Disease**" *Journal of Bio-Science and Bio-Technology* Vol.6, No.2 (2017). https://doi.org/10.14257/ijbsbt.2014.6.2.07
10. Tabassum S, Mamatha Bai B G, Dr. Jharna Majumdar, "**Analysis and Prediction of Chronic Kidney Disease using Data Mining Techniques**", *International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)* ISSN (Online) 2394-2320, Vol 4, Issue 9, September 2017, DOI: 10.13140/RG.2.2.26856.72965.
11. Mamatha Bai B G, Nalini B M, Jharna Majumdar, "**Analysis and Detection of Diabetes using Data Mining Techniques – A Big Data Application in Healthcare"**, *Proceedings of Emerging Research in Computing, Information, Communication and Applications,* Vol 1, *Advances in Intelligent Systems and Computing,* ISBN 978-981-13-5952-1 ISBN 978-981-13-5953-8 (eBook), DOI: org/10.1007/978-981-13-5953-8.
12. Jharna Majumdar, Mamatha Bai B G, Sumant Udandakar, "**Implementation of CURE Clustering Algorithm for Video Summarization and Healthcare applications in Big Data"**, P*roceedings of Emerging Research in Computing, Information, Communication and Applications*, Vol 2, ISBN: 978-981-13-6001-5, DOI: 10.1007/978-981-13-6001-5, *Springer.*
13. Debjani Panda, Ratula Ray, Azian Azamimi Abdullah, Satya Ranjan Dash "**Popular Ensemble Methods: An Empirical Study**" *Journal of Artificial Intelligence Research*, 2017.
14. Shylashree G, Mamatha Bai B G, Jharna Majumdar, "**Statistical Analysis and Prediction of Medical Data using Data Mining Techniques ensuring Data**

Confidentiality", *TOUCAN Research and Development*, ISBN No.: 978-81-942952-8-0, 2019.

15. Madam Chakradar, Alok Aggarwal **"A Machine Learning Based Approach for the Identification of Insulin Resistance with Non-Invasive Parameters using Homa-IR"**, *International Journal of Emerging Trends in Engineering Research*, 8(5), May 2020, 2055 – 2064 https://doi.org/10.30534/ijeter/2020/95852020
16. Edward B. **Panganiban "Microcontroller-based Wearable Blood Pressure Monitoring Device with GPS and SMS Feature through Mobile App",** *International Journal of Emerging Trends in Engineering Research*, 7(6), June 2019, 32-35 https://doi.org/10.30534/ijeter/2019/02762019
17. https://www.kaggle.com/uciml/pima-indians-diabetes-database
18. https://www.kaggle.com/uciml/breast-cancer-wisconsin-data