Data Mining based Diseases Classification

[1]Archika Jain, [2]Devendra Somwanshi, [3]Barkha Narang

[1,2,3]Assistant Professor, [1,3]Department of CSE &[2]EC

[1,2,3]Poornima College of Engineering, Jaipur, India

[1]archikaagarwal@gmail.com, [2]imdev.som@gmail.com, [3]barkhanarang17@gmail.com

Abstract.

Data mining is fast gaining traction in a variety of fields, including organic chemical research, financial forecasting, healthcare, and weather forecasting. Data mining in healthcare is a rapidly growing discipline that may help with prognosis and a better understanding of medical data. Investigation of infirmary for finer well-being decision-making and determent of health centre mistake, early disclosure, and determent of ailment and avertible health centre demise, more utility for funds and cost effectives, and discernment of crooked security claims are all examples of data mining applications in healthcare. Data mining techniques are being used in the detection of a variety of ailments, including diabetes, stroke, cancer, and heart disease. We employ two types of datasets in this study: breast cancer and diabetes databases. We use the WEKA tool to put the techniques into practice. On breast cancer dataset, MLP is better error-free classifier in contrast of remaining with the highest accuracy i.e. 74.12%. On diabetes dataset, SMO is better accurate classifier in contrast of others with the highest accuracy i.e. 79.30%.

Keywords. WEKA tool, Data mining, Diseases.

1. INTRODUCTION

Data mining is an action of analysing enormous data bank to uncover previously unknown patterns, correlations, and information that would be difficult to identify using standard statistical approaches. Data mining is a computational approach that imply the use of intelligent retrieval, predictive analytics, and data bank arrangement to discover figure in enormous number of data file [5]. The extensive motive of the data extract action is to bring out facts from a data cluster and turn it into a formation. Data bank and data administration matter, data pre-processing imitation and inference deliberations, allure measures, difficulty deliberations, post-processing of establish forms, perceptions, and online refurbish are all part of it [3].

In terms of data mining applications, the healthcare business is essential since it generates an extensive range of repository that varies in magnitude, diversity, and rapidness. Condemnatory illnesses such as lymphoma, pulmonary disease, as well as diabetes are among the world's top causes of mortality [6]. Vital information may be obtained from a huge database using data mining tools and methodologies, providing an easy manner for

prophylactic educator to make dominant choices and enhance restorative aid [6]. WEKA is used in view of the fact that it allows ourselves quickly assess and collate knowledge discovery in data algorithms on actual facts [7]. It is now feasible to forecast many disorders more accurately because to developments in computing technology supplied by computer science technologies. As illustrated in Fig 1, data mining may be separated into sub-processes that include data selection, pre-processing, transformation, data mining, and ultimately data interpretation [8]. The classification approach is commonly employed in the health and medical fields. It gives a step-by-step method for creating a classifier model using training data, which is subsequently tested using test data and used to make predictions.

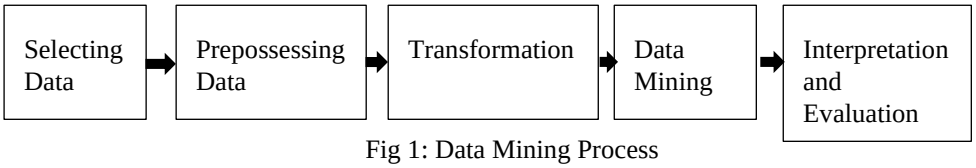

Fig 1: Data Mining Process

*1.1 **Data Excavate Job:*** Fact extraction job may be categorized:

- Prognostic imitation
- Depictive imitation

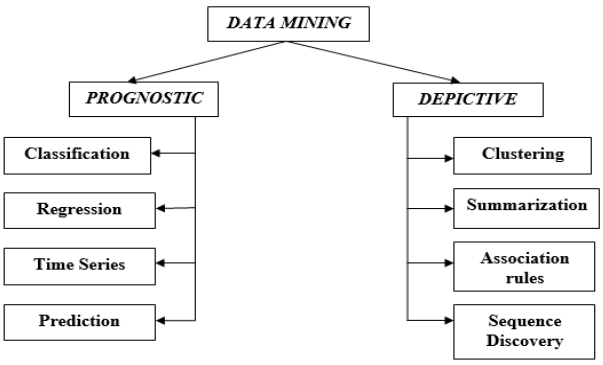

Fig 2: Data Excavate Job

2. CLASSIFICATION IS USED IN PROPOSED WORK

Table 1 List of Attributes

Dataset Name	No. of Attributes	No. of Instances	Attributes
Breast Cancer	9	296	Period, menopause, swell size, bosom canker, vertex
Diabetes	10	788	Pregnant, claret, pressure, skin steroid, paediatric

2.1 Flow Charts of Proposed System:

A flow chart and steps of proposed work as shown:

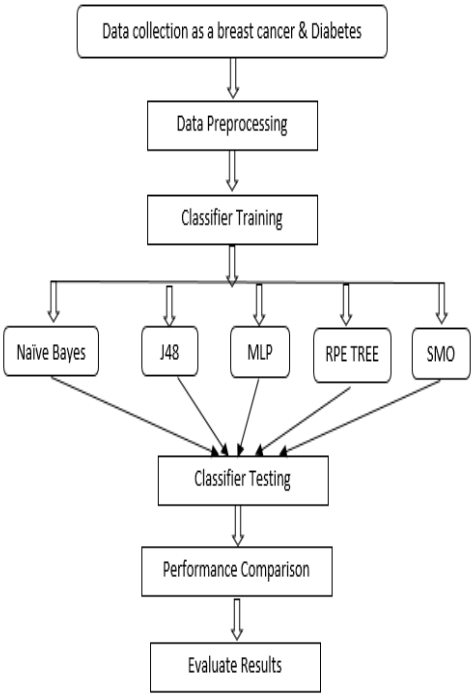

Fig 3: Flow Chart

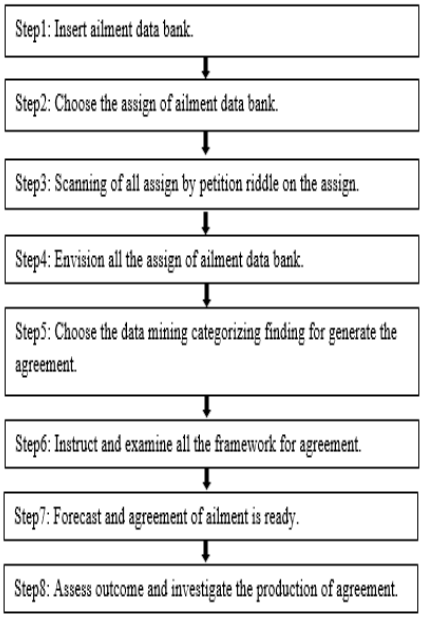

Fig 4: Steps of Proposed Work

3. EXPERIMENTAL WORK

3.1 Breast cancer dataset attributes description:

- Period [5], Menopausal, Quadruple-nodes [3], Fork top, Malignancy [7], Bosom canker [5], Bosom vertex, Betterment, Classification [1].

3.2. Diabetes dataset attributes description:

- Pregnant, Claret, Pressure, Skin, Steroid [3], Multitude, Pediatric [1], Lifetime, Class.

3.3. Performance of classifiers on Breast cancer dataset:

Table 2 Performance of classifiers on Breast cancer dataset

Evaluation Criteria	**Classifiers**				
	Naive Bayes	**J48**	**MLP**	**RPE Tree**	**SMO**
Correctly classified instances	69	66	72	64	68
Incorrectly classified instances	28	31	25	33	29
Accuracy (%)	71.10	68.0	74.12	65.95	70.00

From above table 2 we can conclude that on breast cancer dataset, MLP is better error-free classifier in contrast of others also it is clearly observed meaning it has a higher proportion

of correctly categorized occurrences and a lower proportion of mistakenly classified instances than Naive Bayes, SOM, Rep Tree and J48.

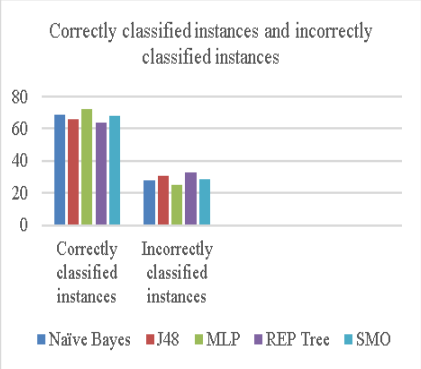

Fig 5: Representational of instances on Breast cancer

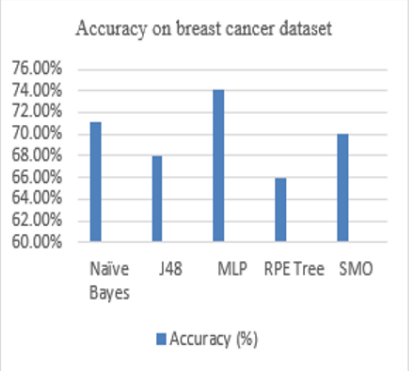

Fig 6: Diagrammatic representation of accuracy on Breast cancer

3.4. Performance of classifiers on Diabetes dataset:

Table 3 Performance of classifiers on Diabetes dataset

Evaluation Criteria	Classifiers				
	Naive Bayes	**J48**	**MLP**	**RPE Tree**	**SMO**
Correctly classified instances	201	199	194	197	207
Incorrectly classified instances	60	62	67	64	54
Accuracy (%)	77.00	76.2	74.30	75.45	79.30

From above table 3, SMO is better error-free classifier and has a higher proportion of correctly categorized occurrences and a lower proportion of mistakenly classified instances than Naive Bayes, MLP, REP Tree and J48.

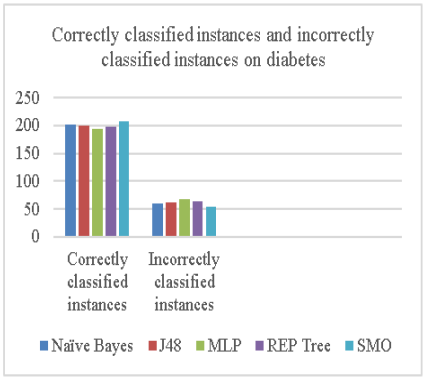

Fig 7: Representational of instances on Diabetes

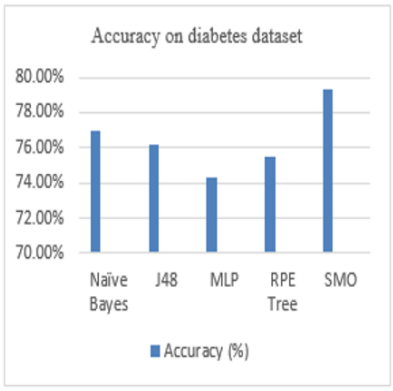

Fig 8: Diagrammatic representation of Accuracy on Diabetes

4. CONCLUSION & FUTURE WORK

We culminate that WEKA instrument is examine as one of the finest apparatus for data extraction categorizing. In this we have used two types of datasets breast Cancer datasets

and Diabetes datasets. And obtain the good results by using the all attributes. On breast cancer dataset, MLP is better error-free classifier in contrast of others also it is clearly observed meaning it has a higher proportion of correctly categorized occurrences and a lower proportion of mistakenly classified instances than Naive Bayes, SOM, Rep Tree and J48. On diabetes dataset, SMO is better error-free classifier in contrast of others also it is clearly observed meaning it has a higher proportion of correctly categorized occurrences and a lower proportion of mistakenly classified instances than Naive Bayes, MLP, REP Tree and J48.

We used percentage split test option for testing the parameters, so in future we will use other testing mode options to increase accuracy of classifiers.

5. REFERENCES

1. S. ALGHUNAIM AND H. H. AL-BAITY, "ON THE SCALABILITY OF MACHINE-LEARNING ALGORITHMS FOR BREAST CANCER PREDICTION IN BIG DATA CONTEXT," IN *IEEE ACCESS*, VOL. 7, PP. 91535-91546, 2019, DOI: 10.1109/ACCESS.2019.2927080.
2. V. MHETRE AND M. NAGAR, "CLASSIFICATION BASED DATA MINING ALGORITHMS TO PREDICT SLOW, AVERAGE AND FAST LEARNERS IN EDUCATIONAL SYSTEM USING WEKA," *2017 INTERNATIONAL CONFERENCE ON COMPUTING METHODOLOGIES AND COMMUNICATION (ICCMC)*, 2017, PP. 475-479, DOI: 10.1109/ICCMC.2017.8282735.
3. N. KUMAR AND S. KHATRI, "IMPLEMENTING WEKA FOR MEDICAL DATA CLASSIFICATION AND EARLY DISEASE PREDICTION," *2017 3RD INTERNATIONAL CONFERENCE ON COMPUTATIONAL INTELLIGENCE & COMMUNICATION TECHNOLOGY (CICT)*, 2017, PP. 1-6, DOI: 10.1109/CIACT.2017.7977277.
4. A. H.J., B. NIRMAL AND A. S. MAHESH, "PROGNOSTICATION OF DIABETES USING DATA MINING MODELS," *2021 6TH INTERNATIONAL CONFERENCE ON COMMUNICATION AND ELECTRONICS SYSTEMS (ICCES)*, 2021, PP. 1883-1887, DOI: 10.1109/ICCES51350.2021.9489061.
5. R. AL-DHAIBANI, M. A. M. BAMATRAF AND K. Q. SHA'AFAL, "DATA BENCHMARK COLLECTION OF PATIENTS WITH MALARIA FOR MACHINE LEARNING: A STUDY IN HADHRAMOUT- YEMEN," *2019 FIRST INTERNATIONAL CONFERENCE OF INTELLIGENT COMPUTING AND ENGINEERING (ICOICE)*, 2019, PP. 1-3, DOI: 10.1109/ICOICE48418.2019.9035166.
6. N. RAMKUMAR, S. PRAKASH, S. A. KUMAR AND K. SANGEETHA, "PREDICTION OF LIVER CANCER USING CONDITIONAL PROBABILITY BAYES THEOREM," *2017 INTERNATIONAL CONFERENCE ON COMPUTER COMMUNICATION AND INFORMATICS (ICCCI)*, 2017, PP. 1-5, DOI: 10.1109/ICCCI.2017.8117752.
7. S. S. RAYKAR AND V. N. SHET, "COGNITIVE ANALYSIS OF DATA MINING TOOLS APPLICATION IN HEALTH CARE SERVICES," *2020 INTERNATIONAL CONFERENCE ON EMERGING TRENDS IN INFORMATION TECHNOLOGY AND ENGINEERING (IC-ETITE)*, 2020, PP. 1-7, DOI: 10.1109/IC-ETITE47903.2020.442.
8. R. SYED, R. K. GUPTA AND N. PATHIK, "AN ADVANCE TREE ADAPTIVE DATA CLASSIFICATION FOR THE DIABETES DISEASE PREDICTION," *2018 INTERNATIONAL CONFERENCE ON RECENT INNOVATIONS IN ELECTRICAL, ELECTRONICS & COMMUNICATION ENGINEERING (ICRIEECE)*, 2018, PP. 1793-1798, DOI: 10.1109/ICRIEECE44171.2018.9009180.

Biographies

Archika Jain received the bachelor's degree in information technology from YIT, Jaipur in 2013, the master's degree in computer engineering from PU, Jaipur in 2016, and pursuing PhD in Computer Engineering. Her research areas include image processing, data mining, deep learning, and machine learning. She has published many research papers in different National and International journals and conferences.

Mr. Devendra Kumar Somwanshi completed his Bachelor’s degree in EC in 2007 from Govt. Eng. College, Bikaner. He did his Master’s Degree in EC from Thapar University, Patiala in 2009 and is currently pursuing his PhD. He authored 3 technical books. He has published more than 60 research papers in National and International conferences.

The Author, Barkha Narang received Bachelor’s Degree in Computer Science Engineering from Rajasthan University in 2004, Masters Degree from Banasthali University in 2009. She is pursuing Ph. D. She has been working on Block chain. She has an experience of 16 plus years into Academics. Also, she has written many research papers.