A Study on Fertility Data using Data Mining Techniques

S. Hemalatha[1] G. Thulasi[2]
[1]Assistant Professor
[1,2]Shrimathi Indira Gandhi College for Women, India

Abstract— An integral part of human beings is reproduction, which is dependent on fertility rates. The focus of the study here is to study the male fertility rates and classify them into a decision making process. Use the following J48, Random Forest and LAD algorithms are used for classification of data sets. The datasets are classified on the basis of the accuracy, error rates like RMSE, MAE etc. and finally predictions are done. The model predicts the fertility class for the input test data. Thus this is a helpful and useful tool in the health industry which will make accurate predictions by increasing prediction rates and helps in making preventive decisions. This enables persons to be easily identified well in advance with good fertility rates or having other defects. Thus the study using data mining techniques helps in the identification of the dataset population to predict the fertility rates and find which method is more accurate or suitable for predicting the fertility of the population in general with the available data.

Keywords: Fertility, Decision Tree. Data Mining for Fertility

I. INTRODUCTION

Infertility refers to the inability for childbearing after at least one-year of marriage for both male and female without the use of external contraception. Both the Infertility and individual face social problems as consequences which are vital issues for couples across the world. Gender wise the cause of male infertility is detectable partially that too only 40% and it is not clinically detectable in nearly 60% of the cases. Also the main reason is that the treatment of infertility in men is much more difficult and complicated than in women according to medics. It has been found out that in the last two decades alone fertility rate has considerably declined more in men than women which suggests that apart from environmental factors other factors like wrong habits lead to such problems like quality of the sperm. Scientific methods can be taken to reduce its negative impacts on both the genders. It has been found that the lifestyle impact on general health includes fertility has become a major area of research. This is due to many factors in the new life style like marriage age, weight, eating habits, lack of exercise, work or peer pressures, mental stresses, pollution and environmental conditions and occupational hazards etc. etc., having a significant impact on fertility rates. Social habits like smoking, drugs and illegal drugs, alcohol and coffee also tend to have a significant negative impact on fertility capacity of the persons doing such acts. The concept of knowledge extraction from data has been gradually used in clinical area.

The knowledge has the slow growth in the health field. Nowadays, the use of artificial intelligence techniques in decision support systems is increasing in the field of medicine from day to day. One of the most important data mining techniques is the classification method. Most important purpose of the classification to obtain a model for the prediction. Diagnosis with laboratory approach includes expensive and sometimes disturbing tests for patients. Therefore, the use of classification that requires only filling out a questionnaire, sometimes can be the first step in deciding whether or not to perform the experimental method.

II. LITERATURE REVIEW

Hadigheh Kazemijaliseh, et al in their work "Prevalence and Causes of Primary Infertility in Iran: A Population-Based Study" investigated the prevalence and causes of primary infertility in Iran. They conducted the study in an urban area of Iran comprising a total of 1067 married women and they were all randomly selected using systematic random sampling method with Unmarried women and also those with unwilling pregnancy during marriage below were excluded. Logistic regression analysis was employed to find primary infertility and this showed that was fertility was independently related to the following factors like old age of women (OR: 1.37; 95% CI: 1.14–13.63, Value: 0.001), higher BMI (OR: 1.95; 95% CI: 1.87–4.14, Value: 0.003), active smoking (OR: 1.47; 95% CI: 1.38–3.53, Value: 0.012) and higher educational level (OR: 2.23; 95% CI: 1.12–5.53, Value: 0.03). Another aspect to consider is Iran having higher primary infertility when compared with the worldwide trends of infertility, thereby indicating that such risks will help healthcare providers to formulate polices to overcome this problem. Hadigheh Kazemijaliseh et al in their work "The Prevalence and Causes of Primary Infertility in Iran: A Population-Based Study" investigated the prevalence and causes of primary infertility based on a study in an urban area of Iran. They considered a total of 1067 married women who participated in the Tehran Lipid and Glucose Study were randomly selected using systematic random sampling and collected data by using validated ad-hoc questionnaires. They found that Iran had high infertility than the other places worldwide thus guiding healthcare providers and policy makers to design and implement interventions to slow down this trend. Faranak Mohammadpour Lashkari et al in their seminal work "Clinical aspects of 49 infertile males with 45,X/46,XY mosaicism karyotype: A case series" recorded data such as height, male general appearance, testis size and volume, external genitalia, spermogram and hormonal levels, testis pathology, Y chromosome microdeletion and karyotype, and assisted reproductive technology (ART) which is the outcome based on patients profile and history. They investigated sixty four infertile males with forty five with X and forty six males with XY mosaicism and found that fifteen cases had structural abnormalities in Y chromosome who were subsequently excluded. Taking approximately 49 available spermogram, 21 cases reported as azoospermic men, while 28 of them classified as nonazoospermic patients in which four of them displayed normal spermogram.Thus the the hormonal tests indicate that there is no tellable

differences between azoospermic and nonazoospermic groups as azoospermia, only three couples underwent an ART cycle in which all of them failed and those From 14 nonazoospermic cases who entered into the ART cycle only three cases experienced successful pregnancy out of which one was twins, thus helping to find occurrence of rare cases. David Gil et al in their work "Predicting seminal quality with artificial intelligence methods" have studied and showed show that Multilayer Perceptron and Support Vector Machines show the maximum accuracy having values of over 87% for several parameters. They collected data by a normalized questionnaire from young healthy volunteers and then, and then use the results of a semen analysis to assess the accuracy in the prediction of the three classification methods mentioned above which is quite to decision trees where visual approaches are shown which are readily understandable and can compensate the lower accuracy levels. This leads to the fact that artificial intelligence methods provide useful tools for prediction using environmental factors and life habits for fertility rates.

Rakesh K Sharma et al in their paper titled "Lifestyle factors and reproductive health: Taking control of your fertility" have made it evident that lifestyle factors have a significant impact on fertility. Lifestyle factors are the modifiable habits and ways of life that can greatly influence overall health and well-being, including fertility and also lifestyle factors like age at which they start a family, good nutrition, weight, exercise, psychological stress, environmental and occupational exposures, and others can have substantial effects on fertility; lifestyle factors such as cigarette smoking, illicit drug use, and alcohol and excessive caffeine consumption can negatively influence fertility rates while other habits may be beneficial for prevention. Jose L. Girela in their work "Semen Parameters Can Be Predicted from Environmental Factors and Lifestyle Using Artificial Intelligence Methods" presented a methodology with useful tools for early diagnosis of patients with seminal (infertile) disorders and also help in choosing candidates for semen donation. They propose Decision Support System by using artificial intelligence in for predicting the semen characteristics.

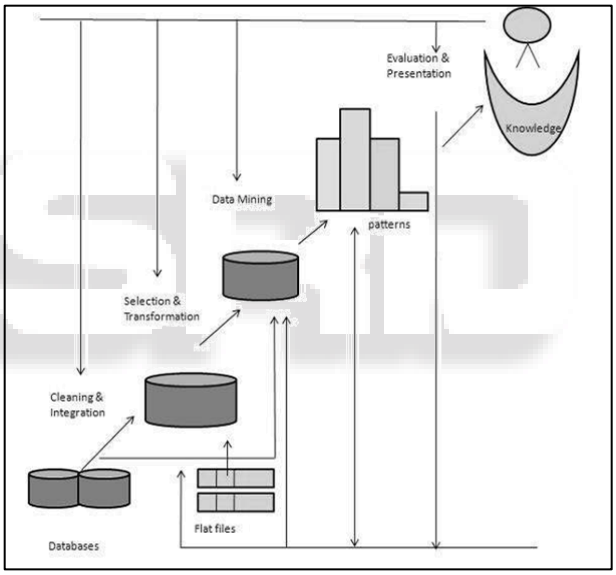
A. *Problem Identification*

There are a lot of studies to study other diseases whereas there are comparatively very few studies, datasets when compared with other disease like fertility, thyroid heart diseases etc. But most models follow black box models – non disclosure of algorithms and methods, when it comes to fertility as this is a social problem or rather a stigma associated with honor and societal imbalances. So disclosure rates are very high in fertility rates. The currently used methods to diagnose fertility rates are physically expensive. Even if a method allows instances to be accurately assigned to groups, no information is provided to users regarding the reasoning underlying that assignment. The models are unpredictable and hence suffers from accuracy. Requires high number of positive samples Involves lots of training errors. The trade-off between margin maximization and error minimization. This leads to lots of false positives.

B. *Problem Statement*

The fertility dataset is a medical dataset used from the UCI repository. The dataset comprising 100 oersins of data comprises both discrete and continuous, i.e., mixed, attributes, which provide a semen sample analyzed according to the WHO 2010 criteria. Sperm concentration are related to socio-demographic data, environmental factors, health status, and life habits The model uses Recursive Rule Generation in combination with J48, Random Forest and LAD The rules generated exhibit good performance, reduced number of rules and relevant input variables. This makes the prediction process accurate – Diagnosis as normal or altered. Rule extraction studies have the capability of providing good explanations. (White Box Model) The extracted rules have a high level of accuracy, particularly in the medical setting, Yet they are simple and easy to understand The prediction model has less false positives and false negatives. The accuracy of the extracted rules for diagnosing thyroid diseases from this dataset are very high.

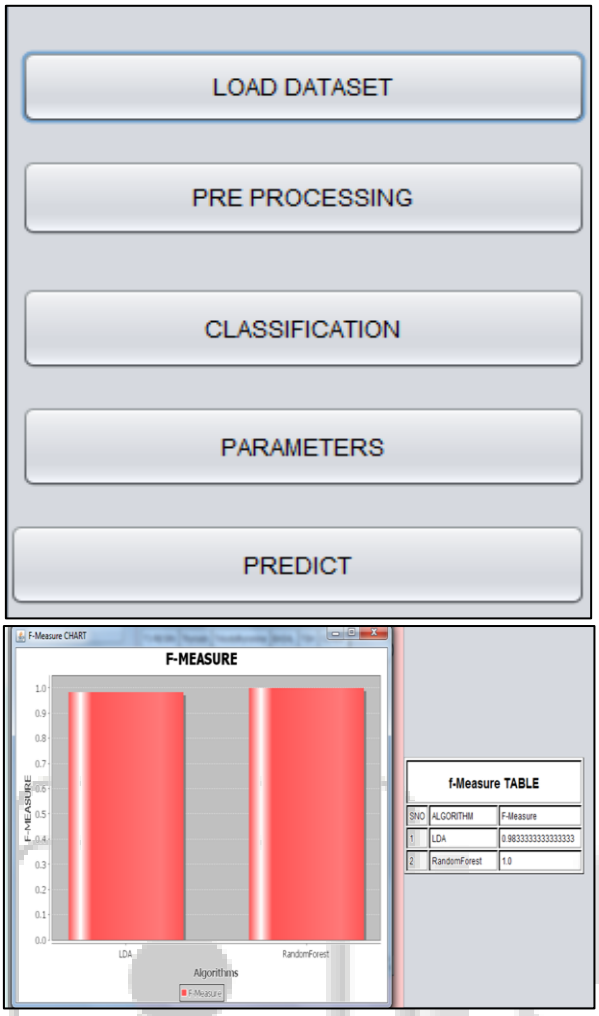
III. ARCHITECTURE MODEL



Such values can be determined by regression, inference-based tools using Bayesian formalism, decision trees, clustering algorithms. Use a decision tree to try and predict the probable value in the missing attribute, according to other attributes in the data. ID3 (Iterative Dichotomiser 3) is an algorithm invented by used to generate a decision tree from a dataset. ID3 is the precursor to the C4.5 algorithms, and is typically used in the machine learning and natural language processing domains.

IV. PSEUDO CODE STEPS

1) Calculate the entropy of every attribute using the data set {\displaystyle S}
2) Split the set {\displaystyle S} into subsets using the attribute for which the resulting entropy (after splitting) is minimum (or, equivalently, information gain is maximum)
3) Make a decision tree node containing that attribute
4) Recurse on subsets using remaining attributes.



V. RESULTS AND DISCUSSION

Generally, for each of the three decision tree algorithms, (i) different values result in different classification accuracies; (ii) there is a value where the corresponding classification accuracy of the Decision Tree is the best; and (iii) the values, in which the best classification accuracies are obtained, are different for both the different data sets and the different classification algorithms.

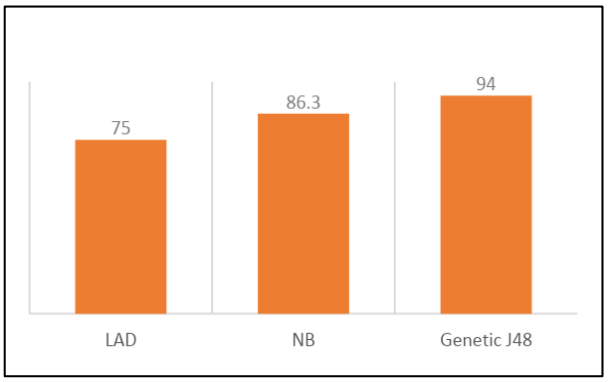
Hence a more appropriate value is desired for a specific classification problem and a given classification algorithm. In most cases, the default values recommended are not the optimal. Especially, in a few cases the corresponding classification accuracies are very small. It has been found that of the three decision tree classification algorithms, (i) different values result in different classification accuracies; (ii) there is a value where corresponding classification accuracy; and (iii) the values, in which the best classification features are got, are different for both the data sets, the modified – genetic J48 Decision Tree model is found to be the best. The results and findings are tabulated below with appropriate charts

ALGORITHMS	LAD	NB	Genetic J48
FEATURES	80.4	86.3	90
EFFICIENCY	90.4	94.8	97.2
ACCURACY	89.6	98.6	95.8

Table 1: Parameters for all three algorithms

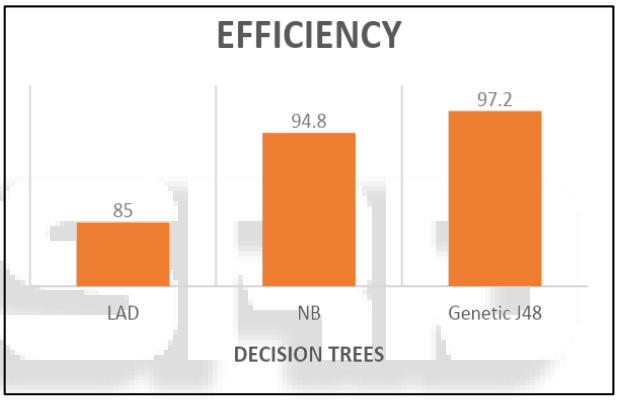
In terms of features Genetic J48 Tree shows the maximum efficiency with 94% followed by NB tree and finally LAD with 75%.

A. *The Features Generation*



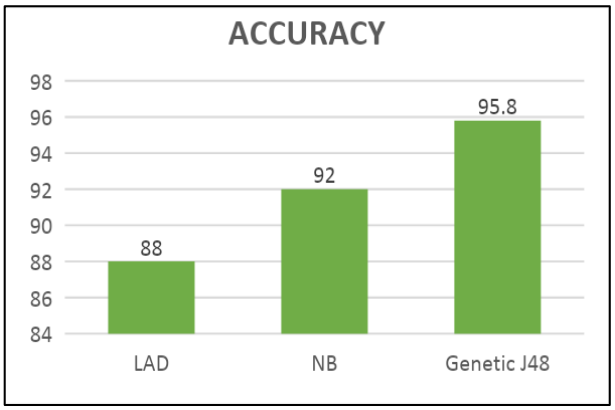
In terms of efficiency again Genetic J48 Tree shows the maximum efficiency with 97.2% followed by NB tree with 94.8% and finally LAD with 85%.

B. *Efficiency of the Algorithms*



Finally when it comes to accuracy the genetic J48 is most accurate as it comes with 95.8% while NB has 92% and LAD Tree has 88% accuracy.

C. *Accuracy of the Algorithms*



This means the results are the best, and the performance is optimal for the genetic J48 tree. For each of the three decision tree algorithms, although the values where the best classification accuracies are obtained are different for various parameter in the dataset, the genetic J48 is the preferred model because the classification accuracies are the best among the lot. When determining the value, besides

classification accuracy, the proportions of the selected features are taken into account.

VI. Conclusion

The proposed genetic model based rule mining and its derived decision tree based application is very fast and accurate. It removes irrelevant features or rules and has got the capacity to handle high dimensional datasets and also does not buckle under dimensionality curse. Each cluster is treated as a single feature and thus dimensionality is drastically reduced. The study thus successfully shows the comparison of the three decision tree classification models for the UCI repository fertility dataset and shows the tree structure formed enabling users to take accurate decisions based on the input parameters. The three algorithms like LAD, NB and J48-Genetically modified are used for testing and classifying the nodes in the tree. Given the fertility data set and its attributes, the ideal scenario would be to have a given set of criteria to choose a proper decision tree algorithm to apply. Choosing a decision model, however, can be a difficult task. Even ending just the most relevant approaches for a given data set is hard. Most of the algorithms generally assume some implicit structure in the data set. One of the most important elements is the nature of the data and the nature of the desired cluster. Another issue to keep in mind is the kind of input and tools that the algorithm requires. This report has a proposal of a new hierarchical clustering algorithm based on the overlap rate for cluster merging. The genetic J48 algorithm provides a natural way of distinguishing clusters and implementing the basic requirement of clustering as high within-cluster similarity and between-cluster dissimilarity. So for scenario based training the algorithms may be used. Further the genetic J48 model is found to be the most efficient and accurate when compared with the other two decision models in terms of time, accuracy and features. In future the models may include other decision support systems with parameters from clinical tests aiding prediction of the fertility. This model will help in the identification of fertility using the genetic model in a very fast and accurate manner. Even the overheads and costs are significantly low. This will ensure a better decision making ability.

References

[1] World Health Organization (WHO). "Infertility: A tabulation of available data on prevalence of primary and secondary fertility," WHO program on maternal and child health and family planning, Division of family health, Geneva, 1991
[2] S. Bhasin, D.M. de Kretser, H.W. Baker, "Clinical review 64: Pathophysiology and natural history of male infertility," J Clin Endocrinol Metab, Vol. 79, pp. 1525-1529, 1994.
[3] D. Gil, J.L. Girela, J. De Juan, M.J. Gomez-Torres, M. Johnsson, "Predicting seminal quality with artificial intelligence methods," Expert Systems with Applications, vol. 39, pp. 12564–12573, 2012.
[4] R. Sharma, K.R. Biedenharn, J.M. Fedor, A. Agarwal,"Lifestyle factors and reproductive health: taking control of your fertility," Reproduction Biology and Endocrinology, vol. 11, 2013, Available at: http://www.rbej.com/content/11/1/66
[5] M.A. Anwar, N. Ahmed, "Analyzing lifestyle and environmental factors on semen fertility using association rule mining," Information and Knowledge Management, vol. 3,pp. 15-21, 2014.
[6] P.S. Duggal, S. Paul, P. Tiwari, "Analytics for the quality of fertility data using Particle Swarm Optimization," International Journal of Bio-Science and Bio-Technology, vol. 7, pp. 39–50, 2015.
[7] M. Naeem, U. Lumière, L. France, "Etiological evaluation of seminal traits using Bayesian Belief Network," International Journal of Bio-Science and Bio-Technology, vol. 6, pp. 79–86, 2014.
[8] R. Battiti, "Using mutual information for selecting features in supervised neural net learning," IEEE Transaction on Neural Networks, vol. 5, pp. 537-550, 1991.
[9] J. Ham, M. Kamber, J. Pei, "Data mining: concepts and techniques," 3rd ed., Elsevier Ltd, pp.378-387 and 418-425, 2000.
[10] Logistic Regression, Available at: http://www.holehouse. org/ mlclass/ 06_Logistic_Regression.html
[11] J.A. Hanley, B.J. McNeil, "The meaning and use of the area under a receiver operating characteristic (ROC) curve," Radiology, vol. 143, pp. 29-36, 1982.
[12] D. Tomar, S. Agarwal, "Feature selection based Least Square Twin Support Vector Machine for diagnosis of heart disease," International Journal of Bio-Science and Bio-Technology, vol.6, pp. 69-82, 2014.

