Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X
IMPACT FACTOR: 5.258

IJCSMC, Vol. 5, Issue. 8, August 2016, pg.5 – 9

DATA MINING TO ELICIT PREDOMINANT FACTORS CAUSING INFERTILITY IN WOMEN

Ms. N. Vijayalakshmi[1], Ms. M. UmaMaheswari[2]

1. Asst. Professor, Dept. of M.C.A., Shrimati Indira Gandhi College, Trichy-2.
2. Research Scholar in Computer Science, Shrimati Indira Gandhi College, Trichy-2.

1. nvijimca@gmail.com 2. umamanohar89@gmail.com

Abstract- One of the most significant issues faced by women these days is infertility. Although several factors are considered to lead to infertility, it would be worth enough to find the most predominant factors causing this problem so that a better and quick solution could be sought in time. Data mining and statistical analysis go hand in hand in identifying these factors from a clinical database containing primary data pertaining to significant factors relating to fertility/infertility in woman. The sample population encompasses both fertile and infertile women relating to a good age spread. Data mining techniques like association rule mining, classification using decision tree induction, clustering, prediction using a decision tree approach and building an application based on the knowledge gained for predicting the probability of infertility in a woman have been used to thoroughly attain our objectives.

Keywords- Data mining, classification, prediction, association rules, statistical analysis, clustering

I. INTRODUCTION

In many countries infertility refers to a couple that has failed to conceive after 12 months of regular sexual intercourse without the use of contraception. Infertility could be primary or secondary. The consequences of infertility are manifold and can include societal repercussions and personal suffering. Causes of infertility may be sexually transmitted diseases, DNA damage, history of diabetes mellitus, thyroid disorders, celiac, adrenal disease, pituitary factors, etc. Environmental factors like consumption of toxins, tobacco may also cause infertility. Common causes of infertility in females include ovulation problems, tubal blockage, pelvic inflammations due to tuberculosis, age-related factors, uterine problems, tubal ligations, endometriosis or advanced maternal age.

This paper uses data mining and statistical analysis techniques to identify the predominant factors causing infertility in women. Already factors that are thought to be significant like age, BMI, primary or secondary infertility, employment, residential area, marital status, history of TB / DM, hormonal levels especially for Follicle Stimulating Hormone(FSH), Lieutinizing Hormone (LH) and Thyroid Stimulating Hormone(TSH) are only considered. Among these the most significant ones leading to infertility are identified. Characteristics of each significant factor is studied in fertile and infertile women leading to knowledge discovery of causes of infertility in each case. The entire data set is also subject to classification using two

different decision tree induction methods and a comparative study of the methods is also undertaken. An attempt is also made to predict infertility in a woman using the knowledge gained through decision tree induction and this is used to build a software model for the same. Association rules that govern infertility are also generated using Association Rule Mining. Clustering is used to perform descriptive data mining.

II. LITERATURE REVIEW

Age, BMI and TSH play a significant role in causing infertility in women.[1] Rising obesity rates present a global public health challenge. Approximately 1.6 billion adults worldwide are overweight (BMI 25-30 kg/m^2) and atleast 400 million were obese (BMI > 30 kg/m^2) in 2005.[2] It is increasing being recognized that obesity also contributes to infertility. Obesity in women has been shown to increase time to conception.[3-5] The relative risk of anovulatory infertility is 2.7 (95% CI, 2.0-3.7) in women with BMI≥ 32 kg/m^2 at age 18,[6] while in ovulatory but subfertile woman the chance of spontaneous conception decreases by 5%[7] for each unit increase in the BMI.

Obesity is also associated with polycystic ovary syndrome (PCOS) which is a heterogeneous condition characterized by oligo or anovulation, hyperandrogenism, menstrual irregularities and subfertilty.[7,8] Obesity which occurs in 30-75% of women with PCOS increases the magnitude of hormonal and metabolic dysfunction in these women.[9]

Hormones play an important role in the development of reproductive function and in the normal regulation of the menstrual cycle. Disruption of the normal secretion of luteinizing hormone (LH) and follicular stimulating hormone (FSH) in response to pulsatile secretion of gonadotrophin releasing hormone is evidenced in a number of reproductive disorders in women. Traditionally, measurements of prolactin and thyroid stimulating hormone(TSH) are considered important components of the evaluation of women presenting with infertility [9]. Thyroid dysfunction interferes with numerous aspects of reproduction and pregnancy. Several studies have indicated the association of hyperthyroidism or hypothyroidism with an ovulatory cycles, decreased fecundity and increased morbidity during pregnancy.

III. METHODOLOGY

A. Data Sources

There are numerous factors causing infertility in women. A Sample populationconsisting of 575 patients who are getting treated in a reputed endometrial research center in Trichy are taken. Physical, environmental and hormonal factors are taken into account. Data pertaining to 154 fertile women and 421 infertile women were collected for the same attributes. A questionnaire was created consisting of various parameters regarding the factors influencing infertility. Those questionnaires were distributed to the patients who are visiting the centre for weekly/monthly check ups. The response was satisfying. Out of several independent attributes collected from outpatients, it is clear that only some of the factors really play a vital role in causing infertility in women.

B. Statistical Analysis

Weka is a workbench that contains a collection of visualization tools and algorithms for data analysis and predictive modeling, together with graphical user interfaces for easy access to these functions.Weka supports several standard data mining tasks, more specifically, data preprocessing, clustering, classification, regression, visualization, and feature selection. WEKA is a popular data mining tool. It is used to analyze the most significant factors causing infertility. It is also used to perform statistical analysis of each individual attribute.

C. Data Mining

Data Mining may be defined as the composite of techniques employed to detect patterns in large datasets to extract hidden pieces of information. It is a fairly new technique used to discover concealed patterns in the behavior of data. While statisticians have for some time been performing Data Mining manually, recent advances in statistical software, computer power and storage capabilities have enabled us to easily and accurately extract hidden patterns from databases.

1) *Use of classification techniques*

Classification is the most commonly applied data mining technique, which employs a set of pre-classified examples to develop a model that can classify the population of records at large. The data classification process involves learning and classification. In Learning the training data are analyzed by classification algorithm. In classification, test data are used to estimate the accuracy

of the classification rules. If the accuracy is acceptable the rules can be applied to the new data tuples. Decision tree induction is a popular technique used for classification and prediction.

2) Use of k-means clustering

Clustering can be said as identification of similar classes of objects. By using clustering techniques we can further identify dense and sparse regions in object space and can discover overall distribution pattern and correlations among data attributes. Classification approach can also be used for effective means of distinguishing groups or classes of object but it becomes costly so clustering can be used as pre-processing approach for attribute subset selection and classification.

3) Use of association rule mining

Association Rule Mining is a popular and well researched method for discovering interesting relations between variables in large databases. To select interesting rules from the set of all possible rules, constraints on various measures of significance and interest can be used.

IV. RESULTS AND DISCUSSION

A. Use of statistical analysis on the sample data using WEKA revealed the following facts:

1. Patients with age lesser than 28 are more likely to become pregnant than the others.
2. Patients with BMI > 29 are more prone to infertility.
3. Primary infertility is more prevalent than secondary infertility.
4. Patients with a history of diabetes mellitus may be related more to infertility
5. Patients with a history of tuberculosis are highly prone to be infertile.
6. If FSH > 8.5, or LH > 7 or if TSH > 4 the chances of being infertile is very high.

B. Use of CFS Subset Evaluator to identify the most deterministic factors causing infertility produced the following results. BMI, Age, DIABETIC, TB, FSH, LH and TSH are the significant factors leading to infertility (ie., 7 out of 9 factors- residential area, marital status were left out.)

C. Two different classification techniques used produced the following results:

TABLE I
STRATIFIED CROSS VALIDATION SUMMARY OF TWO DIFFERENT CLASSIFICATION TREES

=== Stratified cross-validation === === Summary ===		
	J48 pruned tree technique	**Random Tree**
No. of leaves	14	-
Size of tree	27	459
No. of records / attributes	575 / 9	575 / 9
Correctly Classified Instances	554 96.3478 %	494 85.913%
Incorrectly Classified Instances	21 3.6522 %	81 14.087%
Kappa statistic	0.9074	0.653
Mean absolute error	0.0475	0.1466
Root mean squared error	0.1888	0.3531
Relative absolute error	12.1007 %	37.3463%
Root relative squared error	42.6341 %	79.7435%
Total Number of Instances	575	575

TABLE II
COMPARATIVE STUDY OF TWO DIFFERENT CLASSIFICATION TREES -DETAILED ACCURACY BY CLASS

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
J48 pruned tree	0.942	0.029	0.924	0.942	0.932	0.908	0.958	0.860	FERTILE
	0.971	0.058	0.978	0.971	0.975	0.908	0.958	0.977	INFERTILE
Random Tree	0.792	0.116	0.713	0.792	0.751	0.655	0.857	0.699	FERTILE
	0.884	0.208	0.921	0.884	0.902	0.655	0.857	0.910	INFERTILE

Hence we find that J48 pruned tree is a relatively better technique in terms of accuracy in classifying the given record sets with an accuracy of 96%

D. Use of K-means clustering on the given data set produced the following results:

TABLE III
RESULT ANALYSIS OF K-MEANS CLUSTERING ON THE GIVEN DATA SET

	Initial Cluster centroids		**Final Cluster Centroids (after 3 iterations)**		
	Cluster 0	**Cluster 1**	**Full Data**	**Cluster 0**	**Cluster 1**
	32.5 – 35.2	-inf-21.7	24.4 – 27.1	24.4 – 27.1	24.4 – 27.1
	27.5 – 29.6	25.4 – 27.5	23.3 – 25.4	25.4 – 27.5	23.3 – 25.4
Prim/Sec	Primary	Secondary	Primary	Primary	Primary
Diabetic	No	No	No	Yes	No
Tuberculosis	Positive	Negative	Negative	Negative	Negative
FSH	16.37 – 18.26	8.81 – 10.7	5.03 – 6.92	5.03 – 6.92	5.03 – 6.92
LH	1.729 – 3.338	1.729 – 3.338	4.947 – 6.556	4.947 – 6.556	4.947 – 6.556
TSH	1.81 – 3.41	1.81 – 3.41	1.81 – 3.41	5.01 – 6.61	1.81 – 3.41
Result	Infertile	Fertile	Infertile	Infertile	Fertile

Within cluster sum of squared errors 2405.0

Clustered Instances

0 381 (66%)
1 194 (34%)

E. Use of association rule mining on WEKA for the given data yielded the following results:

Minimum support: 0.2 (115 instances) Minimum metric <confidence>: 0.9
Number of cycles performed: 16

Generated sets of large itemsets:

Size of set of large itemsetsL(1): 16 Size of set of large itemsets L(2): 29
Size of set of large itemsetsL(3): 8

Best rules found:

TABLE IV
ASSOCIATION RULE MINING WITH WEKA ON GIVEN DATASET

	Association rules found	Support	Confidence	Lift	Level	No. of records	Convergence
1	TB=positive ==> RESULT=INFERTILE	171	1	1.37	0.08	45	45.8
2	RESULT=FERTILE ==> TB=negative	154	1	1.42	0.08	45	45.8
3	DIABETIC=no RESULT=FERTILE ==> TB=negative	141	1	1.42	0.07	41	41.93
4	PRISEC=primary TB=positive ==> RESULT=INFERTILE	119	1	1.37	0.06	31	31.87
5	PRISEC=primary RESULT=FERTILE ==> TB=negative	115	1	1.42	0.06	34	34.2
6	DIABETIC=yes 266 ==> RESULT=INFERTILE 253	253 / 266	0.95	1.3	0.1	58	5.09
7	TSH='(5.01-6.61]' 130 ==> RESULT=INFERTILE 123	123 / 130	0.95	1.29	0.05	27	4.35
8	LH='(8.165-9.774]' 140 ==> RESULT=INFERTILE 131	131/ 140	0.94	1.28	0.05	28	3.75

9	PRISEC=primary DIABETIC=yes 181 ==> RESULT=INFERTILE 169	169/181	0.93	1.28	0.06	36	3.73
10	DIABETIC=yes TB=negative 159 ==> RESULT=INFERTILE 146	146/ 159	0.92	1.25	0.05	29	3.04

F. A C program to predict infertility based on the J48 decision tree was tested with the given training set. An accuracy of **86%** was obtained.

V. CONCLUSION

In this paper, we made an attempt to use **data mining as a tool** for analyzing clinical data records of both fertile and infertile patients. Data mining is a powerful tool which is currently used for extracting significant information from historical data. This information can be used for further decision making and prediction. WEKA was used for applying various data mining techniques like Statistical analysis, Associative rule mining. Clustering. Classification and Subset evaluation. This has been very helpful in extracting key information regarding infertility. Two classification methods were used on the same record set to produce almost similar results at varying levels of accuracy. Among them, J48 pruned tree has been found to be more accurate.Clustering is also carried out to verify the output of previous methods. From the information gained an attempt is also made to build a decision tree model for prediction of infertility. The accuracy of prediction is 86%.

ACKNOWLEDGEMENT

We acknowledge with thanks the support given by Ramakrishna Nursing Home and Endometrial Research Centre, Trichy for having provided us with primary clinical data relating to fertility on 575 fertile and infertile women patients only for the purpose of research on infertility.

REFERENCES

1. K.Meena, N.Vijayalakshmi, "Analysis of Factors Causing Infertility in Women using Statistical Analysis and Asssociation Rule Mining", *Indian Journal of Public Health Research & Development,* Vol 6(2), pp 112 – 117, April-June 2015
2. Gesink Law DC, Maclehose RF, Longnecker MP, "Obesity and time to pregnancy", *Human Reproduction,*Vol. 22(2), pp 414-420, Feb 2007
3. Nohr EA, Vaeth M, Rasmussen S, Ramlau-Hansen CH, Olsen J. "Waiting time to pregnancy according to maternal birthweight and prepregnancy BMI", *Human Reproduction,*Vol. 24(1), pp 226–32, Jan 2009
4. Wise LA, Rothman KJ, Mikkelsen EM, Sørensen HT, Riis A, Hatch EE. "An internet-based prospective study of body size and time-to-pregnancy" *Human Reproduction,*Vol. 25(1), pp 253-264, Jan 2010
5. Van der Steeg JW, Steures P, Eijkemans MJ, Habbema JD, Hompes PG, Burggraaff JM, et al. "Obesity affects spontaneous pregnancy chances in subfertile, ovulatory women", *Human Reproduction,* Vol. 23(2), pp 324-328, Feb 2008
6. Metwally M, Li TC, Ledger WL. "The impact of obesity on female reproductive function", *Obesity Reviews,* Vol 8(6), pp 515–23, Nov2007
7. Pasquali R, Gambineri A. "Polycystic ovary syndrome: A multifaceted disease from adolescence to adult age", *Annals of the New York Academy of Science,* Vol 1092, pp 158-174, Dec 2006
8. Diamanti-Kandarakis E. "Role of obesity and adiposity in polycystic ovary syndrome", *International Journal of Obesity,* Vol 31(Supp 2), pp S8–13, Nov 2007
9. G.E., Pontikides, N., Kaltsas, T., Papadopoulou, P., Paunkovic, J., Paunkovic, N. &Duntas, L.H. (1999), "Disturbances of menstruation in hypothyroidism", *Clinical Endocrinology,* Vol 50(5), pp 655–659, May 1999.

