The comparison of decision tree and k-NN to analyze fertility using 2 different filters

By asto buditjahjanto

The comparison of decision tree and k-NN to analyze fertility using 2 different filters

IOP Conf. Series: Materials Science and Engineering **434** (2018) 012048 doi:10.1088/1757-899X/434/1/012048

The comparison of decision tree and k-NN to analyze fertility using 2 different filters

N Rochmawati[1]*, H B Hidayati[3], Y Yamasari[1,4], I G P Asto[1] and L Rakhmawati[2]

[1] Department of Informatics, Engineering Faculty, Universitas Negeri Surabaya, Surabaya, Indonesia
[2] Department of Electrical Engineering, Engineering Faculty, Universitas Negeri Surabaya, Surabaya, Indonesia
[3] Department of Neurology, Medical Faculty, Airlangga University, Surabaya, Indonesia
[4] Department of Electrical Engineering, Engineering Faculty, Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia

*naimrochmawati@unesa.ac.id

Abstract. Fertility is a crucial issue for married couples for ages and a significant clinical problem today. Not only do women have an undue burden of responsibility in fertility regulation but also men. There are some major causes and risk factors for male infertility i.e., environmental factors, life style factors etc. In this paper, we will analyse the infertility using two algorithms, decision tree and k-nearest neighbours. We will do experiments using different splits of training data and different filters. The purpose is to determine which algorithm is more accurate between 2 algorithms either using filter or not. The result shows DT has better performance in accuracy when using dataset without filter and when using randomize filter while k-NN has better performance when using resample filter.

1. Introduction

It is a fact that fertility rate has declined in almost every part of the world, not only in developing countries [1] but also in Denmark [2]. In a news mentioned that the diagnosis causes infertility is more due to men than women. Many factors influence the sperm to be diagnosed as normal or not. The factors are socio-demographic data, environmental factors, health status, and life habits [3].

Many researches about fertility and quality of sperm was carried out using many methods. Seminal quality was Predicted Using Clustering-Based Decision Forests [4], fertility was analysed by using supervised and unsupervised data mining techniques [5], [6], best sperm was selected to increase implantation rate by applying data mining techniques [7], fertility was analysed by using graph based data mining [8], and the last, it is analysed using neural network [3].

Classification technique is a technique where dataset is categorized into number of classes. There are many types of classification algorithms in machine learning. Two of these classification algorithms are decision tree and k-nearest neighbor. The Decision Tree is one of the most popular and most commonly used classification algorithms. While k-nearest neighbor is a simple and powerful algorithm. The aim is to compare both the decision tree and k-nearest neighbor algorithm so that we know which better performance is either when using filter or not.

In this paper we will use both algorithms by comparing them to analyse the fertility dataset. The goal is to find out which algorithm is superior in analysing the fertility dataset either when using filters or no filters.

The experiment has 3 steps: first, we compare 2 algorithms to process raw dataset without filter, second, we compare 2 algorithms to process dataset with randomize filter, and third, we compare 2 algorithms to process dataset with resampling filter.

2. Experiment

This experiment uses weka 3.8.2. and the fertilty data set used was extracted from UCI machine learning repository. This comparison used two different filters: randomize and resampling. There are 100 instances and 10 attributes.

Basically the research conducted has many processes as in figure 1. First we select dataset and open CSV. The file we used is from the website UCI machine learning repository. This website is dedicated to researchers to get data for machine learning and intelligent systems. If the experiment done without filter, then we select directly. But if we use filter for processing data then choose the filter used. Next, process the result. Then the output will be displayed. The experiment is repeated many times because we compare between decision tree and k-nearest neighbor. Each of them is carried out using filter resampling and randomize and no filter.

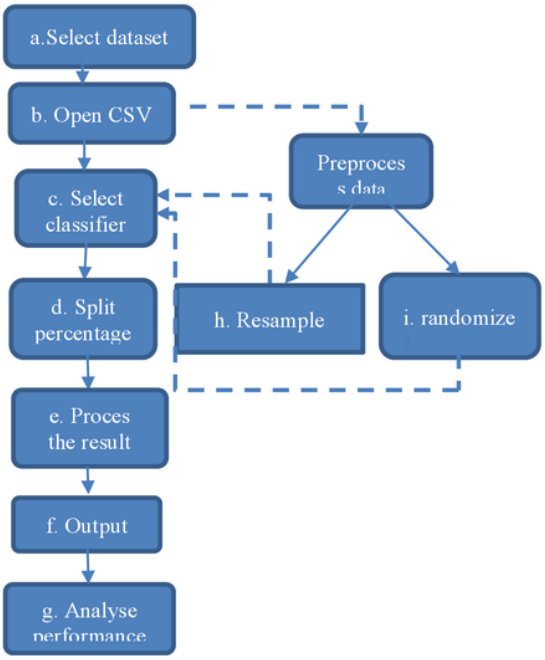

Figure 1. The order steps of experiment.

The experiment was conducted in 3 stages:

a. Compare 2 algorithms without using filter. The step sequence is a, b, c, d, e, f, g as in figure 1.
b. compare 2 algorithms by using randomize filter. The step sequence is a, b, i, c, d, e, f, g as in figure 1
c. compare 2 algorithms by using resampling filter. the step sequence is a, b, i, c, d, e, f, g as in figure 1

The three experimental stages are performed using different splits.

3. Result and discussion

First, the experiment is carried out using no filter. The split percentage using are 50, 60, 70, 80 and 90. When the split percentage using is 50 means that the training data is 50% and the testing data is 50 %. If split percentage is 70 means that the training data is 70% and the testing data is 30%. The result of this experiment is comparing 2 algorithms using false prediction matrix. False prediction has 2 kinds. The first is false positive (FP) and the rest is false negative (FN). False positive means we predicted normal but actually in the fact it is altered. FN is false negative means we predicted altered but actually in the fact it is normal.

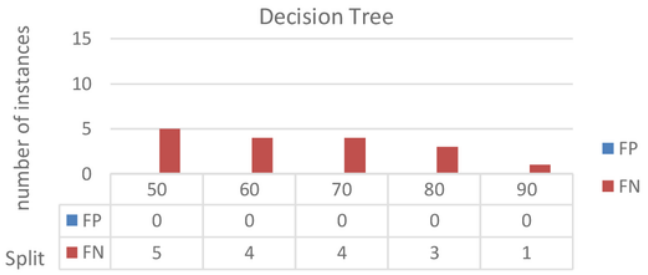

Figure 2. False prediction of decision tree without filter.

Figure 2 shows false prediction of decision tree without filter. The worst result is in split 50 where FN is 5 which is the biggest number of false prediction compared to other split percentage. The best result is in split 90 where FN is 1. It means the false prediction has only 1 false result where we predicted altered but actually in the fact it is normal.

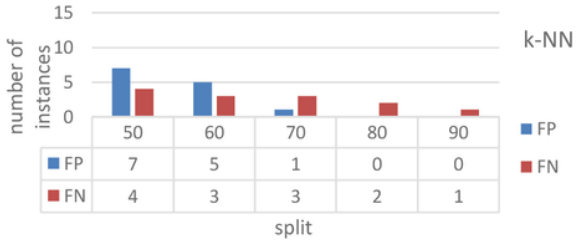

Figure 3. False prediction of k-NN without filter.

Figure 3 shows false prediction of k-nearest neighbor without filter. The worst result is in split 50 where FP is 7 and FN is 4. This is the biggest number of false prediction compared to other split percentage. The best result is in split 90 where FN is only 1. It means the false prediction has only 1 false result.

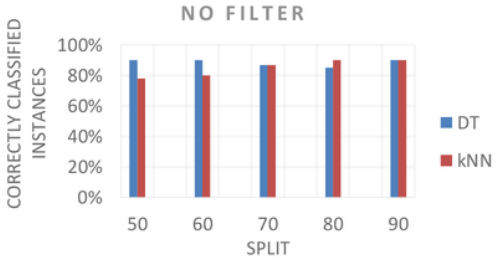

Figure 4. The comparison of accuracy without filter.

Figure 4 shows the result of the comparison between decision tree and k-nearest neighbor where in this case the experiment using no filter. The ordinate number shows correctly classified instances and the

absis number shows the split percentage. Correctly classified instances is resulted from weka application. It means that the higher number correctly classified instanes, the more accurate. The color of blue belongs to decision tree and the orange color belongs to k-nearest neighbor. From figure 4 shows that decision tree is more accurate than k-NN because the result is higher than k-NN in all split except in split 80.

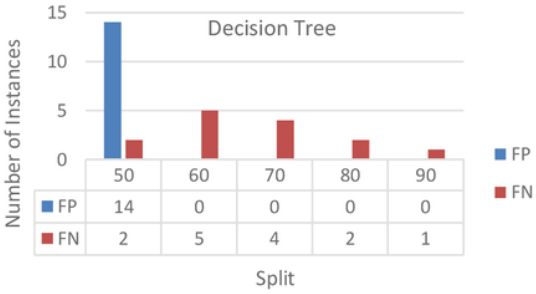

Figure 5. False prediction of decision tree using randomize filter.

Figure 5 shows false prediction of decision tree using randomize filter. The worst result is in split 50 where FP is 14 and FN is 2. this is the biggest number of false prediction compared to other split percentage. The best result is in split 90 where FN is 1.

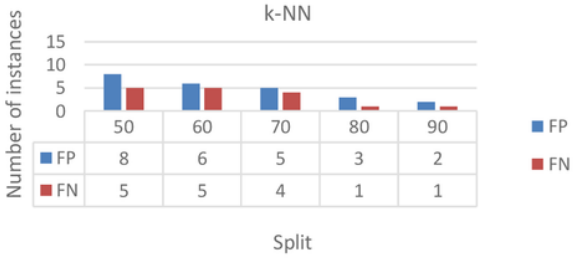

Figure 6. False prediction of k-NN using randomize filter.

Figure 6 shows false prediction of k-nearest neighbor using randomize filter. The worst result is in split 50 where FP is 8 and FN is 5. This is the biggest number of false prediction compared to other split percentage. The best result is in split 90 where FP is 2 and FN is only 1.

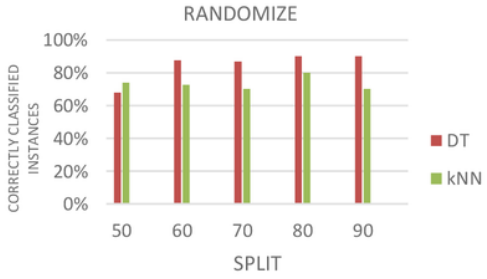

Figure 7. The comparison of accuracy using randomize filter.

Figure 7 shows the result of the comparison between decision tree and k-nearest neighbor where in this case the experiment using randomize filter. The ordinat number shows correctly classified instances and the the absis number shows the split percentage. Correctly classified instances is resulted from weka

application. It means that the higher number correctly classified instanes, the more accurate. The color of orange belongs to decision tree and the grey color belongs to k-nearest neighbor. From figure 7 shows that decision tree is more accurate than k-NN because the result of decision tree is higher than k-NN in all split except in split 50.

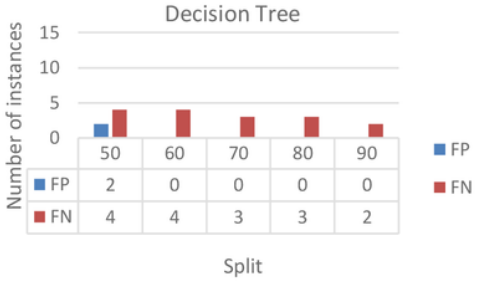

Figure 8. False prediction of decision tree using resampling filter.

Figure 8 shows false prediction of decision tree using resample filter. The worst result is in split 50 where FP is 2 and FN is 4. this is the biggest number of false prediction compared to other split percentage. The best result is in split 90 where FN is 2.

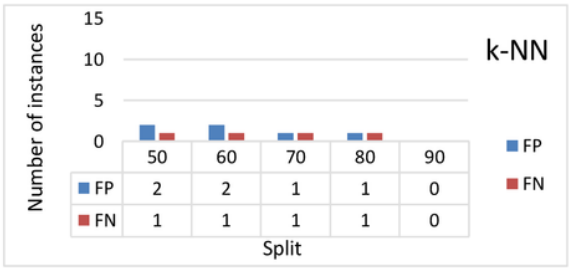

Figure 9. False prediction of k-NN using resampling filter.

Figure 9 shows false prediction of k-nearest neighbor using resample filter. The worst result are in 2 splits, 50 and 60, and where FP is 2 and FN is 1. This is the biggest number of false prediction compared to other split percentage. The best result is in split 90 where there is no false prediction. It means all data is predicted correctly.

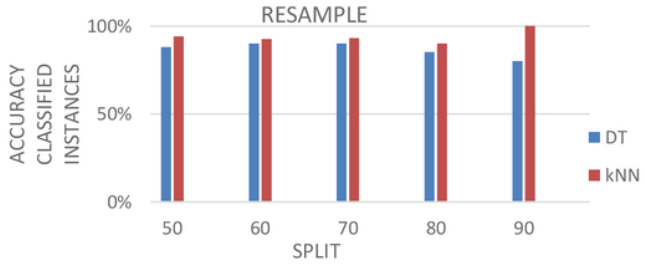

Figure 10. The comparison of accuracy using resampling filter.

Figure 10 shows the result of the comparison between decision tree and k-nearest neighbor where in this case the experiment using randomize filter. The ordinat number shows correctly classified instances and the absis number shows the split percentage. Correctly classified instances is resulted from weka applicaton. It means that the higher number correctly classified instanes, the more accurate. The color

of blue belongs to decision tree and the orange color belongs to k-nearest neighbour. From figure 10 shows that k-NN is more accurate than decision tree because the result of k-NN is higher than decision tree in all split. k-NN has 100% accurate in split 90.

Discussing the result, from figure 2, 3, 5, 6, 8, 9 shows the same pattern. If the training data (percentage split) has small number, then the result is worse than the higher number of split percentage or in other words the greater percentage split (the training data) the better the result. Therefore 90 split percentage has better result than 50 split percentage it was mentioned in some researches [9], [10] that 2/3 of the data sets (67%) for the training set were more robust than 1/2 for the training set (50%). but in other research [11] 1/2 to training set is a little better than the 2/3 rds. to training split. indeed, in 3 studies it was not mentioned that the higher percentage split for training set results more accurate, but only stated that 2/3rd is better than 1/2 and vise versa. Next, from figure 4, 7, 10 it can be concluded that the appropriate algorithm when the data set is not filtered is the decision tree because decision tree has higher accuracy than k-NN. When applied randomize filter then the superior method is the decision tree because decision tree has higher accuracy than k-NN. When the experiment is applied using resample filter then the superior method is k-NN method because the k-NN has higher accuracy than decision tree. It is not about which algorithm is more accurate but which algorithm is more appropriate whether using filter or specified filter. In this case, if the filter implemented is randomize then the appropriate algorithm is decision tree but if the filter is resampling then choose k-NN as the algorithm.

4. Conclusion

In this paper, we will analyse the infertility using two algorithms, decision tree and k-nearest neighbours. We will do experiments using different splits of training data and different filters. The purpose is to determine which algorithm is more accurate between 2 algorithms either using filter or not. The result shows DT has better performance in accuracy when using dataset without filter and when using randomize filter while k-NN has better performance when using resample filter.

Limitations in this study is the dataset used is still unbalance. For future research, before a filter is implemented, data balancing may be necessary. In addition, there should also be comparison using many other datasets to determine whether the results remain consistent as in this paper or not.

References

[1] M Eslami 2016 Decreasing Total Fertility Rate in Developing Countries **10**(4) pp. 163–164
[2] M B Jensen, L Priskorn, T K Jensen and A Juul 2014 Temporal Trends in Fertility Rates: A Nationwide Registry Based Study from 1901 to 2014 pp. 35–39
[3] D Juan, J L Girela, D Gil and M Johnsson 2013 Semen Parameters Can Be Predicted from Environmental Factors and Lifestyle Using Artificial Intelligence Methods 1 **88** pp. 1–8
[4] H Wang, Q Xu and L Zhou 2014 Seminal Quality Prediction Using Clustering-Based Decision Forests pp. 405–417
[5] E Fabio, P Paola, A Jorge and P Melo Fertility 2016 Analysis Method Based on Supervised and Unsupervised Data Mining Techniques **11**(21) pp. 10374–10379
[6] A J Sahoo and Y Kumar 2014 Seminal quality prediction using data mining methods **22** pp. 531–545
[7] S A Mirroshandel, F Ghasemian and S Monji-azad 2016 Applying data mining techniques for increasing implantation rate by selecting best sperms for intra-cytoplasmic sperm injection treatment *Comput. Methods Programs Biomed*
[8] N Bothra and A R Gupta 2014 Graph-based Data Mining : A New Approach for Data Analysis **5**(5) pp. 1230–1236
[9] K K Dobbin and R M Simon 2011 Optimally splitting cases for training and testing high dimensional classifiers *BMC Med. Genomics* **4**(1) p. 31
[10] Rosenwald A, Wright G, Chan W C, Connors J M, Campo E, Fisher R I and Hurt E M 2002 The use of molecular profiling to predict survival after chemotherapy for diffuse large-B-cell lymphoma *New England Journal of Medicine* **346**(25) 1937-1947

[11] Boer J M, Huber W K, Sültmann H, Wilmer F, Von Heydebreck A, Haas S and Vingron M 2001 Identification and classification of differentially expressed genes in renal cell carcinoma by expression profiling on a global human 31 500-element cDNA array *Genome research* **11**(11) 1861-1870

The comparison of decision tree and k-NN to analyze fertility using 2 different filters

ORIGINALITY REPORT

18%

SIMILARITY INDEX

MATCH ALL SOURCES (ONLY SELECTED SOURCE PRINTED)

★N Saurina, E Wahyuningtyas, T S Rini, S W Purwaningrum, F S Rejeki. "Information system to measure healthy home", IOP Conference Series: Materials Science and Engineering, 2018 — 8%

Crossref

EXCLUDE QUOTES	OFF	EXCLUDE MATCHES	OFF
EXCLUDE BIBLIOGRAPHY	OFF		