EasyChair Preprint

№ 8855

Using Decision Tree and Naive Bayes to Predict Kidney Stones Disease

Samer Nofal and Rana Nidal

September 22, 2022

Using Decision Tree and Naive Bayes to predict kidney stones disease

[S]amer Nofal
Department of Computer Science
German Jordanian University
Amman, Jordan
Samer.nofal@gju.edu.jo

[R]ana Nidal Abo Orouq
Department of Software Engineering
Princess Sumaya University for Technology
Amman, Jordan
ran20198016@std.psut.edu.jo

***Abstract*—High incidence of diseases related to kidney stones disease become one of the major concerns in health care systems around the world. Early Prediction of this disease decreases its appearance and its related costs, using classification which is one of the data mining techniques that can help in making the best predictions. This study aimed to develop a model for the early detection of kidney stones to provide a decision support system. The data was collected from 500 patients who visited the urology clinic in Rosary's sister's hospital in Irbid from 2017 through 2019, and some of them were diagnosed to have kidney stones whereas the others were diagnosed with different diseases. The gathered data was analyzed using the WEKA toolkit, which provides many data mining algorithms such as Decision Tree J48 and Naïve Bayes, which were used in this paper to build a predictive model. The results show the effectiveness of the model build using the Naïve Bayes algorithm for predicting a kidney stone disease. Moreover, according to the applied models, show the family history of kidney stones, is the most vital parameter in the prediction of kidney stones disease.**

***Index Terms*—classification, data mining, Decision Tree, Naïve Bayes, kidney stone.**

I. Introduction

The huge amount of data nowadays leads to difficulties for humans to do data mining manually, data mining is the way of discovering and extract knowledge and information from huge datasets. While (Naik and Samant, 2016) [1] defines Data mining as "the process of discovering interesting knowledge from large amounts of data stored in databases, data warehouses, or other information repositories". The appearance of automated data mining tools become a must according to the explosion of data, and using such tools assist by reducing the time-consuming in extract useful knowledge from the available data through using machine learning algorithms [1]. Those tools are used in different fields such as in healthcare systems, collect data from patients is an important step, leads to a large volume of data that need to analyze through applying Machine learning algorithms offered in data mining tools to identify the important parameters to predict diseases in a short time and high precision. In this paper, Decision Tree J48 and Naïve Bayes were used to build a classification model to predict a patient with kidney stones disease. In the following sections, a brief background about Machine Learning definition and Types will be displayed in the next section. Related work conducted to predict a kidney stones disease, will be reviewed in the third section. Then, the methodology used to build a classification model will be displayed in detail in the fourth section. While in the fifth section a comparison between the two classifiers builds using Decision Tree J48 and Naïve Bayes. In section six the results will be discussed and analyzed. Finally, a conclusion will appear in the last section to conclude the work conducted in this paper.

II. Background

A. Machine learning Definition

Machine Learning (ML) is a form of artificial intelligence which is a form of computer science with software capable of self-modification programs capable of improving themselves, ML is used to interpret or extract information from data and allow machines to deal with data more efficiently, to learn from the data. The use of ML has become a necessity due to an abundance of data sets in various fields such as education, military, medicine, and many others [3].

B. Types of Machine Learning

Two types of machine learning techniques have been discussed below.

1) Supervised learning: In this type machine learning algorithms need human assistance to provide a dataset, then the dataset will be divided into two data sets training and testing data set. Training data set used to build the model by learning some pattern, then applied over testing data set to make prediction or classification. Here are the three most popular supervised Machine Learning algorithms [3].

- Decision Tree
  Sorting attributes to groups according to their values, it is a tree of rules consist of a root node, branches, and leaf node which describes a decision. Where nodes represent attributes, branches represent the possible value of the node [2]. As illustrated below in Fig.1. It is mostly used for classification.

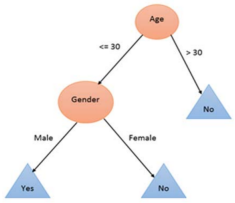


Fig. 1. Example of Decision Tree [2]

- Naïve Bayes

  Generates trees based on the likelihood of conditional probability occurring. As illustrated in Fig.2. These trees are also called the Bayesian network [3]. It is used for clustering and classification.

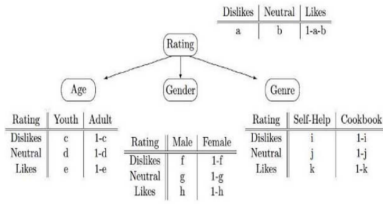


Fig. 2. Example of Naïve Bayes [3]

- Support Vector Machine (SVM)

  It is a state-of-the-art machine learning technique, in general, it depends on margin calculation, by draw margins between the classes with maximum distance. It is mostly used for Classification.

  2) Unsupervised Learning: The unsupervised learning algorithms discovered some features from the data, then use these learned features to determine the class of the new data entered. No need for external assistance and it is used for clustering. Two main algorithms used are discussed below.

- K-Means Clustering
  It creates groups automatically when the items that have similar features are put together in the same group or cluster [4]. As illustrated below in Fig.3.

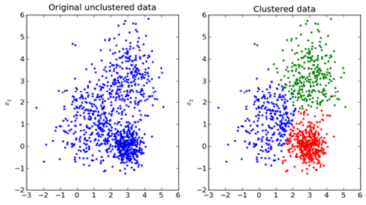


Fig. 3. Example of K-Means Clustering [4]

- Principal Component Analysis (PCA)

  Focus on decreasing dimensions of the data, to make calculations easier and faster [5]. Fig.4. shows an example of how PCA works, two- dimensional data (2D) after applying PCA over the data, it will be 1D.

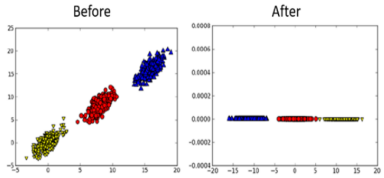


Fig. 4. Example of applying PCA over data [5]

The previous two types of Machine Learning, Supervised Learning and Unsupervised Learning are the most common. The following Figure 5 mentioned the other types of Machine Learning.

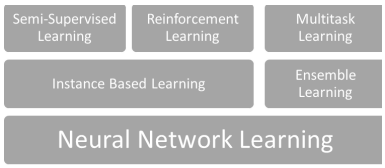


Fig. 5. Types of Machine Learning

III. Literature Review

Many studies have been done for the prediction of kidney diseases such as kidney stones, using data mining algorithms to create a predictive model, based on existing data. Here some of these studies. Kumar and Abhishek [5] build a model to diagnose kidney stone disease, three types of neural network algorithms, learning vector quantization (LVQ), multilayer perceptron (MLP), and radial basis function (RBF) were applied to examine the most effective one due to their accuracy, time consumed to build a model and the size of training data set. According to the studied algorithms, the best model build using MLP gained the highest accuracy 92% and decreasing the diagnosis time. In [6] study aimed to build an effective discriminant model, two classification methods were applied discriminant analysis (DA), and artificial neural networks (ANNs) over genetic and environmental factors such as water, outdoor activities, and milk consumption. When using genetic variables and environmental factors ANN model was better with 89% accuracy, where DA was classifying 74% of participants successfully. When using only the genetic factors both models revealed almost the same results. Moreover, (Kazemi and Mirroshandel, 2018) [7] various methods such as Bayesian, Decision Trees, Artificial Neural Networks, and Rule-based classifiers, were used to create a predictive model for early detection of the kind of kidney stone and to determine the effective parameters. Also,

four models generated using ensemble learning to enhance the accuracy of every learning algorithm then applied a novel technique for merging single classifiers in ensemble learning, final ensemble learned model was powerful with high accuracy of 97.1% could be applied over future cases to predict the chance of nephrolithiasis occurrence. Regarding the previous models, the following parameters such as sex, acid uric condition, calcium level, hypertension, diabetes, nausea and vomiting, flank pain, and urinary tract infection (UTI) were identified as the most essential parameters for the prediction of nephrolithiasis.

Regarding the previous researches about kidney stones disease, show the importance of building a predictive to early detection, and diagnosing this disease. The researchers in this study used different supervised learning algorithms and comparing between them to build an effective predictive model, the method used in this study are discussed in detail in the following section.

IV. Methodology

A. Patients

The researchers collected the data used in this study by visited Rosary Sisters Hospital in Irbid. Data comprising of 500 patients who visited Dr. Firas Sahawnes' Urology clinic, from 2017 through 2019. The patients were complaining of various urological symptoms such as left flank pain, hematuria, right flank pain, urinary frequency, suprapubic pain, bilateral flank pain, dysuria, urinary incontinence, and intermittency. Some of the patients were diagnosed to have kidney stones, whereas the others did not have.

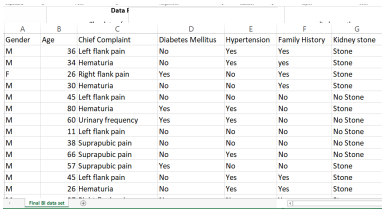
B. Data Set Features

There are many risk factors or parameters such as gender, age, diabetes mellitus, hypertension, and family history of kidney stones, which are considered as the most vital parameters for predicting the chance of kidney stones. Therefore, the prepared dataset considered these factors as features. These features and their possible values are described below:

- Gender: this feature has two values male and female.
- Age: this feature displays the age of patients in years.
- Chief complaint: this feature displays 9 urological symptoms were patients complaining from. These symptoms are: left flank pain, hematuria, right flank pain, urinary frequency, suprapubic pain, bilateral flank pain, dysuria, Urinary incontinence, and intermittency.
- Diabetes Mellitus: in this feature two values were shown, Yes refers to patients complaining from diabetes, where No refers to patients that not suffering from diabetes.
- Hypertension: two possible values for this feature, Yes refers to patients with Hypertension and No refers to patients who do not have Hypertension.
- Family history: when patients have a family history of kidney stones the value of this feature will be Yes, otherwise No.
- Kidney stone: this is the class (the feature that the classification model will predict), when patients suffering from kidney stone the value is Stone, otherwise No Stone.

C. Data Preparation

The data of patients were collected as word files. A set of preparation was applied over the data to discover the data set need to build the model, after extract and reviewed the data, the researchers identify the common and frequently data that the doctor asked about from each patient, which mean these data are important to decide whether the patient has a kidney stone or not, then these vital parameters become a feature of the data set as mentioned in the previous section. These features and their values were presented using an Excel sheet as illustrated below in Fig.6. After all, the generated Excel sheet was converted to (CSV) format, finally converted to (arff) format to become suitable to deal with through the WEKA data mining tool.



Gender	Age	Chief Complaint	Diabetes Mellitus	Hypertension	Family History	Kidney stone
M	36	Left flank pain	No	Yes	Yes	Stone
M	34	Hematuria	No	Yes	yes	Stone
F	26	Right flank pain	Yes	No	Yes	Stone
M	30	Hematuria	No	No	Yes	Stone
M	45	Left flank pain	No	No	No	No Stone
M	80	Hematuria	Yes	Yes	No	Stone
M	60	Urinary frequency	Yes	Yes	No	No Stone
M	11	Left flank pain	No	No	No	No Stone
M	38	Suprapubic pain	No	No	No	No Stone
M	66	Suprapubic pain	No	Yes	No	No Stone
M	57	Suprapubic pain	Yes	No	No	Stone
M	45	Left flank pain	No	Yes	Yes	Stone
M	26	Hematuria	No	Yes	Yes	Stone

Fig. 6. Example of a data set

D. Method

Many data mining tools are available such as Rapid miner, WEKA and Knime, etc. the researchers chose one of the most widely used toolkit WEKA (Waikato Environment for Knowledge Analysis), it is open source and easy to use also it provides a large set of Machine Learning algorithms, that used for various data mining techniques such as classification, association rules, and clustering. It allows data pre-processing and results in visualization. In this study two data classification algorithms, Decision tree J48 and Naïve Bayes have been applied to build a classification model.

V. Comparison

To examine the accuracy of each model the following evaluation techniques have been used, the 10-folds cross-validation, percentage split with 66%, and training and testing data sets were applied.

When using decision tree J48 with 10-folds cross-validation applied the model was correctly classified 469 instances out of 500 instances, with 93.8% accuracy of the model as shown in Fig.7. While the percentage split has been applied the model was correctly classified 152 instances out of 170 with 89.4118% accuracy, as shown in Fig.8.

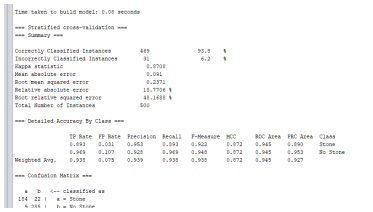


Fig. 7. Accuracy of 10-folds cross validation through Decision Tree J48

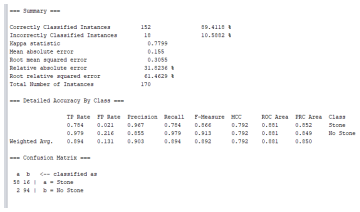


Fig. 8. Accuracy of percentage split through Decision Tree J48

When training and testing data sets were applied at the same time, the model has correctly classified 94 instances out of 100 patients, with 94% accuracy. As illustrated in Fig.9. To split the data set into training and testing data sets, the researchers used Resample, which is an unsupervised WEKA filter. It is divided the instances of the data set provided randomly into training and testing data sets without redundancy of instances that means the instances chose in the training data set are not included in the testing data set.

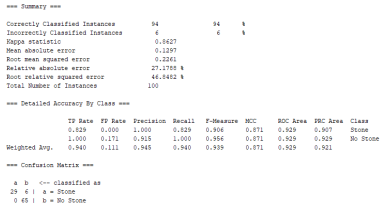


Fig. 9. Accuracy of the model while training and testing data set applied together through Decision Tree J48

While using the 10-folds cross-validation in Naïve Bayes, the model was correctly classified 469 instances out of 500 patients with 93.8% accuracy, this result is like Decision Tree J48 results through using cross-validation. Moreover, when percentage split was applied the model has correctly classified 160 instances out of 170 patients, with 94.1176% accuracy, but when training and testing data sets were applied at the same time the accuracy of the model was 95%, which is a high percentage of accuracy.

TABLE I
ACCURACY OF THE MODEL WHILE TRAINING AND TESTING DATA SET APPLIED TOGETHER THROUGH DECISION TREE J48

Algorithm	10-folds cross validation	66% percentage split	Supplied test set
Decision Tree J48	93.8%	89.4118%	94%
Naïve Bayes	93.8%	94.1176%	95%

According to the previous results and the comparison of accuracy produced by applying a Decision Tree J48 and Naïve Bayes algorithms, displayed in Table 1, the model built using Naïve Bayes, has higher accuracy than the model built using Decision Tree J48, in all evaluation techniques. This shows the effectiveness of the model build using the Naïve Bayes algorithm for predicting a kidney stones disease.

VI. Results Analysis

The results found that the Family history of kidney stone was the most vital parameter in identifying kidney stone. In this study, almost all the patients with a family history of kidney stones were complaining of a kidney stone, also regarding the results, males are more prone to kidney stones than females. Patients with hypertension and/or diabetes mellitus are more prone to kidney stones than patients not suffering from these diseases. The most common symptoms that indicate kidney stones were Left flank pain, right flank pain, and hematuria. These parameters are considered to be the most vital parameters that could affect the prediction of kidney stones disease.

VII. Conclusion

The widespread of kidney stones among people all over the world, motivate the researchers to find a predictive model helping in early diagnosing kidney stones patients to determine the appropriate treatment and decrease the effects of this disease. A predictive model was established as a result of this study to predict a kidney stones disease, this classifier could be used to classify new patients in the future, with high accuracy and without requiring expensive experiments to predict the kidney stone. This research was conducted according to data collected from patients of Rosary's sister's hospital in Irbid.

For future research, data can be gathered from different hospitals in various cities and/or countries, to create a more generalized and effective prediction model.

Acknowledgment

The authors would like to acknowledge the cooperation of urologist Dr. Firas Sahawne.

References

[1] Naik, A. and Samant, L. (2016). Correlation Review of Classification Algorithm Using Data Mining Tool: WEKA, Rapidminer, Tanagra, Orange and Knime. Procedia Computer Science, 85, pp.662-668.
[2] L. Rokach, O. Maimon, "Top – Down Induction of Decision Trees Classifiers – A Survey", IEEE Transactions on Systems
[3] A. Singh, N. Thakur and A. Sharma, "A review of supervised machine learning algorithms," 2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom), New Delhi, 2016, pp. 1310-1315.
[4] http://pypr.sourceforge.net/kmeans.html
[5] Kumar K, Abhishek B. Artificial neural networks for diagnosis of kidney stones disease; 2012.
[6] Chiang D, et al. Prediction of stone disease by discriminant analysis and artificial neural networks in genetic polymorphisms: a new method. BJU Int 2003;91(7):661–6.

[7] Kazemi, Y. and Mirroshandel, S. (2018). A novel method for predicting kidney stone type using ensemble learning. Artificial Intelligence in Medicine, 84, pp.117-126.