EARLY PREDICTION AND DIAGNOSIS OF CHRONIC KIDNEY DISEASE (CKD) USING WEKA TOOL AND APRIORI ALGORITHM

Akshita Sharma¹, Bhanu Pratap Singh², Lakshya Garg³, Dr. Anil Kumar K. ⁴

¹Student, SCOPE, VIT University
²Student, SCOPE, VIT University
³Student, SCOPE, VIT University
⁴Associate Professor, SCOPE, VIT University

ABSTRACT

Over many years we have seen massive changes in how technology has been affecting our lives. Technical advancements in web and data technology have been around for quite a lot of time now and during all this time we have realized how they have been changing different sectors of human life. Healthcare industry has seen a new dawn upon it since it has been touched by these technical advancements, earlier what was considered as obscure and irrelevant has now been given a meaning all thanks to web and data technology. Due to the advent of data mining we have been able to use the large piling data of patients and records which hold sufficient information to find pattern and develop prototypes. The already proven fact that healthcare industry upholds voluminous data which can be explored to help the people with chronic diseases such as Chronic Kidney Disease (CKD). In this paper we tend to analyze the attributes, which play an important role in deducing whether a person is suffering from CKD. For this purpose we implement Apriori algorithm with the help of WEKA tool to execute a real time dataset and identify patterns that give us a much clearer idea about the indicated problem. We also use the classification techniques such as Naïve-Bayes classifiers, Decision Tree and clustering techniques i.e. K-means clustering to create a better understanding of the situation and get inferences from it.

Keywords - Classification, Clustering, Attributes, TP rate, F-measure, Precision.

[1] INTRODUCTION

Healthcare is arguably amongst the most important sectors in all the fields of knowledge and study, in recent years many advancements have been made in the healthcare sector in the surge to make the human life better. Besides these advancements there is always a scope of bringing up a method to create something better. One such area that offers for massive growth for both human knowledge and life expectancy is chronic diseases. Chronic diseases refer to those diseases which lasts with a person over a long period of time. One such chronic disease is the Chronic Kidney Disease (CKD).
The kidneys are a very important part of the human body. They function in a manner to keep the human body fit and clean, they filter the blood and also eliminate the wasteful material in the body and are also responsible for the formation of urine and balancing fluids in the body. Chronic Kidney Disease is a far more extensive than what is known by the people. Due to its lack of understanding it quite frequently goes unfound and undiagnosed in its primary stages. CKD is a peculiar disorder in which the kidneys tend to reduce its functionality and size moreover losing the capability to remove the wasteful material from the body which leads to high amount of wastes in the body which over an extended period of time makes the body prone to diseases. It is very natural for people to ignore the fact that they have CKD at reduction of kidney function as low as 25 percent of the standard. As the disease becomes more and more dangerous for the patient it acts critically and causes waste to build up inside the host body. The treatment of CKD is focused upon slackening and blocking the effect of the disease.

[2] RELATED WORK

Diagnosis of Chronic Kidney Disease (CKD) by Using effective classification and feature selection techniques:

In this paper various classification methods accuracy was tested on a chronic kidney disease dataset which was taken from a UCI repository. To compare the classification done by different classifiers weka tool was used and accuracies for different classifiers was calculated.

Analysis of the effectiveness of the Apriori Algorithm in the application of Medical Billing using Data Mining:

This research shows associations between diagnosis and treatment. The same kind of algorithm has been used as is used in simple billing, although some minor modifications have been made.

Implementation of WEKA tool for medical data classification for early CKD prediction:

Writers have thought about J48 along with Naïve Bayes, SVM based, k-NN classifiers and Random Forest utilizing execution estimates like measurements of kappa, RMSE, ROC and MAE utilizing Waikato (WEKA tool) apparatus. Creators likewise thought about these particular classifiers on different exactness estimates like True Positive rate, False Positive rate, accuracy, review and the likes of f-measure by actualizing on the tool (WEKA). Trial inferences demonstrates the fact that irregular woods classifiers have the better order exactness over others for incessant kidney sickness related data logs.

CKD Prediction on the basis of Imbalanced Data using Multilayer Perceptron:

This paper centers around class lopsidedness in restorative information, when building up a classifier for basic leadership in endless kidney infection. The examination appears, testing calculations can enhance the execution of arrangement calculations and the rate of learning is an imperative value which altogether impacts on multiple layers of perceptron.

Using Association Analysis to Find Diseases Related to Childhood Obesity:
Explores related infections to stoutness y in youngsters and youths through affiliation examination. This paper indicates what illnesses occur in youth stoutness.

A NOVEL IMPUTATION METHOD FOR EFFECTIVE PREDICTION OF CORONARY KIDNEY DISEASE:

This paper portrays an answer for handle missing qualities for interminable kidney illness dataset. Proposed WAELI calculation predicts the missing qualities utilizing single esteem attribution, numerous esteem ascription and the last esteem is anticipated by registering weighted normal of each model.

Un-Apriori: a Novel Association Rule Mining Algorithm for Unstructured EMRs:

In this paper, the investigation of the issue of fascinating affiliation administer mining on unstructured EMR information is demonstrated. First attributes of unstructured EMR information is examined by investigating genuine in-quiet office EMR of interminable malady , and afterward set straight with an advanced novel affiliation control mining calculation, called Un-Apriori, in light of Apriori. Un-apriori is an enhanced calculation for mining unstructured data.Un-Apriori embraces an ordered preprocess conspire for EMR information to fulfill prerequisites of Apriori calculation.

Analysis of a Population of Diabetic Patients Databases in Weka Tool:

The examination is meant to discover the characterters that decide the persistance of diabetes and find the highest number of people dealing with diabetes alongwith 249 populace utilizing Waikato(WEKA tool) apparatus. In this paper the information order is patients dealing with diabetes informational collection is created by gathering information from doctor's facility archive comprises of 249 occurrences with 7 unique characteristics.

Mining the Efficient Association Rules using Apriori Algorithm through Comparative Analysis and Attribute based analysis of the different types of Association Rule structures:

In this paper we utilize apriori calculation to discover the affiliation rules, considering the 3 affiliation administer calculations i.e. apriori , prescient apriori and tertius .

Data Mining Applications In Healthcare Sector:

A Study An assortment of procedures have been engaged and analyzed, distinctive instruments and, methodologies and its effect on social insurance. This paper means to make a point by point examine results of different sorts of information extraction applications in the human services part and to decrease the multifaceted nature of the investigation of the social insurance information exchanges.

Comparison of Different Classification Techniques Using WEKA for Breast Cancer:

This paper means to break down the execution of various techniques for characterization and clustering for an arrangement of extensive information. The calculations tried are Radial Basis Function, Pruned Tree, Single Conjunctive Rule Learner and NN Algorithm, Bayes Network.

Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction:
In this paper the problem of compelling and abridging diverse calculations of information mining Utilized in the field of restorative forecasts are broke down. The fundamental spotlight is on utilizing distinctive calculations and mixes of a few target traits for shrewd and powerful heart assault expectation utilizing information mining.

Analysis of WEKA Data Mining Algorithm REPTree, Simple Cart and Random Tree for Classification of Indian News:

This paper has been done for assessment of execution of REPTree, Simple Cart and Random Tree arrangement calculation with regards to dataset of Indian news. Subsequently it is discovered that Random Tree calculation performs best in Arranging all the News.

Analysis of Clustering Algorithm of Weka Tool on Air Pollution Dataset:

In this paper, calculations and strategies are assessing and differentiating the different clustering calculation by utilizing WEKA device to discover which calculation will be more agreeable for the clients for performing clustering calculation.

Analysis of road traffic fatal accidents using data mining techniques:

In this paper the creator applies measurements examination and information mining calculations on the FARS Fatal Accident dataset as an endeavor to address the issue of mischances because of street activity.

[3] BACKGROUND

Diagnosis

Some of the basic tests are performed to test if the patient has symptoms for CKD:

Blood test - decide if waste substances are as a rule sufficiently sifted through.
Urine test - see if there is either blood or protein in the urinal stream.
Kidney scans - decide if there are any blockages in the urinal stream, uncover state of kidney.
Biopsy of Kidney - little example of the tissues presiding in kidney is pulled back after being tried regarding the damages that are incurred by the cell.
Chest X-ray – Major objective is regarding to test for the liquid that is held in the lungs(pneumonic edema).
Glomerular filtration rate (GFR) - thinks about the levels of waste products

Symptoms

CKD is a gradually progressive disease. It can happen at a point as low as 25 percent of the regular function. Even after one kidney stops functioning other carries on with it’s work, this disease poses a real threat as it is only detectable in its advanced stages. Regular checkups are mandatory for people having kidney related issues.

[4] DATA MINING

Data mining refers to creation of useful information from the raw data. Companies utilize software to identify models and prototypes in sizeable batches of extractable data, and use it so as to get more information about their target customers. It is very useful since a business can acquire a lot of knowledge if they tend to learn about the thinking of their customers and then accordingly make efficient marketing
techniques and upsurge sales whilst reducing costs. The concept of Data mining revolves around efficient
collection of information and warehousing along with processing. Usage of data mining in the healthcare
industry occurred in the middle of 1990’s, since then it has worked it’s way up as a revolution in the
healthcare industry, reason being that enormous volume of data which earlier had not been processed and
mined are collected, all this information contains several relationships and patterns which can distinctively
help in prediction and detection of the disease in many different fields of medicine. This can lead to an all
important reduction in the mortality rate and help sustain life of infected people.

Classification

Classification is the function of data mining whose job is to allocate the item set in a compendium to specific
classes and categories. This is done by prediction of the consequence by analyzing the input set through the
algorithm as it checks the set of attributes and their corresponding outputs.

Clustering

Clustering is an unsupervised learning technique used for combining objects of a particular set by analyzing
their characteristics, and arranging them in regard with their similarities.

Association Rules

Mining of association rules are a very impactful aspect in regards to datamining, it presents a process that
is intended to find frequent patterns, association or fundamental arrangement from information found in
different data repositories.

[5] METHODOLOGY

In this experiment we have used weka tool to analyze our data , this data which we have taken from a
hospital is being analyzed using apriori algorithm to predict which attributes in the following data will
probably lead to a chronic kidney disease . We have used various classifying clustering and associating
algorithms to determine the probability of people having ckd. The primary data set used consists of a total
of 150 instances which have been compared to a secondary data set consisting of 400 instances taken from
UCI machine learning repository. Our data sets contains 7 attributes.

<table>
<thead>
<tr>
<th>S.no</th>
<th>Attribute Used</th>
<th>Type of Attribute</th>
<th>Attribute Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>age</td>
<td>Numeric</td>
<td>Age</td>
</tr>
<tr>
<td>2.</td>
<td>sex</td>
<td>Nominal</td>
<td>Sex</td>
</tr>
<tr>
<td>3.</td>
<td>hb</td>
<td>Numeric</td>
<td>Hemoglobin</td>
</tr>
<tr>
<td>4.</td>
<td>blu</td>
<td>Numeric</td>
<td>Blood Urea</td>
</tr>
<tr>
<td>5.</td>
<td>blg</td>
<td>Numeric</td>
<td>Blood Glucose</td>
</tr>
<tr>
<td>6.</td>
<td>sc</td>
<td>Numeric</td>
<td>Serum Cretanine</td>
</tr>
<tr>
<td>7.</td>
<td>ckd</td>
<td>Nominal</td>
<td>Chronic Kidney Disease</td>
</tr>
</tbody>
</table>
The following clustering models have been used in our experiment:

1. **Naive Bayes**: These come in the family of probabilistic classifiers and were quite helpful as our main aim was to calculate the probability. These being highly scalable can be used for calculating a lot of parameters i.e. big datasets.

<table>
<thead>
<tr>
<th>Class</th>
<th>Rate of True Positive</th>
<th>Rate of False Positive</th>
<th>Precision</th>
<th>Recall</th>
<th>F-Measure</th>
<th>MCC</th>
<th>ROC Area</th>
<th>PRC Area</th>
</tr>
</thead>
<tbody>
<tr>
<td>y</td>
<td>0.597</td>
<td>0.446</td>
<td>0.588</td>
<td>0.597</td>
<td>0.603</td>
<td>0.603</td>
<td>0.173</td>
<td>0.599</td>
</tr>
<tr>
<td>n</td>
<td>0.554</td>
<td>0.382</td>
<td>0.566</td>
<td>0.554</td>
<td>0.569</td>
<td>0.573</td>
<td>0.173</td>
<td>0.599</td>
</tr>
<tr>
<td>Avg.</td>
<td>0.587</td>
<td>0.414</td>
<td>0.587</td>
<td>0.587</td>
<td>0.586</td>
<td>0.590</td>
<td>0.173</td>
<td>0.599</td>
</tr>
</tbody>
</table>

*Table 1.2 Performance measure of Naïve Bayes Classifier*
**Random Forest:** Random Decisions forests is a method wherein multiple learners are trained to solve a single problem, used for classification and regression, these use multiple number of decision trees at training time.

**Random Tree:** Random Tree refers to randomly built trees derived by stochastic approach. It is built on a random subset of columns.
Table 1.3 Performance measure of Random Tree

Simple K-means: This is a simple unsupervised learning algorithm. Consists of a straightforward and simple approach to group a given informational index through a specific number of bunches settled.

Performance Measures Used:

1. Accuracy

Accuracy is ascertained as the total number of two right predictions (TP + TN) partitioned by the total number of a dataset (P + N).

Accuracy: Eqn 1.1 used in Table 1.2,1.3

\[
\text{Accuracy} = \frac{(TP + TN)}{(TP + TN + FP + FN)}
\]
2. Precision

The quantity related to the positively classified were significant. A test can cheat and augment this by just returning positive on one outcome it's most certain about.

Precision : Eqn 1.2 used in Table 1.2,1.3

\[
\frac{TP}{(TP + FP)}
\]

3. Recall

The measure of how great a test is at recognizing the positives. A test can cheat and expand this by continually returning "positive".

Recall : Eqn 1.3 used in Table 1.2,1.3

\[
\frac{TP}{(TP + FN)}
\]

[5] EXPERIMENTAL RESULTS

Following is the graphical representation of all the attributes in their

The above diagram depicts variation of the attribute values in their original format.
Following is the graphical representations of all the attributes after all of them have been converted to nominal attributes and are discretized so as to be eligible for applying Apriori algorithm.

This diagram depicts the variation of attribute values when they are converted to nominal values.

Apriori
The bar graphs represent the variation of all the performance measures applied on 3 different types of classifiers.

[6] SUMMARY

In this paper, we explored what are the probable causes of chronic kidney diseases, these provide information and vital ramifications to the patients looking to enhance their health and do the treatment. This study demonstrated that besides the statistical strategies or clinical trials utilized as a part of the healthcare field there is an alternative method, alternative way such as data mining technique which can be applied to large amount of data such as those collected in the healthcare industry.

In any case, this study is focused on a particular disease as it were. In this manner, Future studies are expected to yield more significant outcome when analysed with successive example investigations of disease and different variables that decide disease criteria.
REFERENCES


