DOI: 10.24996/ijs.2022.63.1.33

ISSN: 0067-2904

CART_based Approach for Discovering Emerging Patterns in Iraqi Biochemical Dataset

Sarah Sameer*[1], Suhad Faisal Behadili[1], Mustafa S. Abd[1], Ali Salam[2]
[1]Department of Computer Science, College of Science, University of Baghdad, Baghdad, Iraq
[2] Earthlink Company, Baghdad, Iraq

Received: 30/12/2020 Accepted: 9/3/2021

Abstract

This paper is intended to apply data mining techniques for real Iraqi biochemical dataset to discover hidden patterns within tests relationships. It is worth noting that preprocessing steps take remarkable efforts to handle this type of data, since it is pure data set with so many null values reaching a ratio of 94.8%, then it becomes 0% after achieving these steps. However, in order to apply Classification And Regression Tree (CART) algorithm, several tests were assumed as classes, because of the dataset was unlabeled. Which then enabled discovery of patterns of tests relationships, that consequently, extends its impact on patients' health, since it will assist in determining test values by performing only relevant tests. Therefore decreases the number of tests for patients.

Keywords: CART, Data mining, Biochemical, Iraq.

لأكتشاف الانماط الناشئة في البيانات البيوكيمائية العراقيةCART النهج القائم على خوارزمية

ساره سمير*[1]، سهاد فيصل[1] ، مصطفى عبد[1] ، علي سلام[2]
[1]قسم الحاسبات، كلية العلوم، جامعة بغداد، بغداد، العراق
[2]شركة ايرثلنك ، بغداد، العراق

الخلاصة

تهدف هذه الورقة إلى تطبيق تقنيات التنقيب عن البيانات لمجموعة بيانات كيميائية حيوية عراقية حقيقية لاكتشاف الأنماط المخفية ضمن علاقات التحاليل البيوكيميائية. ومن الجدير بالذكر أن خطوات المعالجة المسبقة تتطلب جهودًا ملحوظة للتعامل مع هذا النوع من البيانات ، نظرًا لأن مجموعة البيانات النقية يوجد بها العديد من القيم الفارغة تصل الى نسبة 94.8% والتي تصبح 0% بعد تطبيق هذه الخطوات. ومن اجل تطبيق خوارزمية شجرة التصنيف والانحدار تم افتراض العديد من التحاليل كفئات بسبب كون مجموعة البيانات غير مصنفة. والتي ساعدت في كشف أنماط علاقات التحاليل والذي بالتالي يوسع تأثيرها على صحة المرضى. حيث ستساعد في تحديد قيم التحاليل من خلال اجراء التحاليل المرتبطة بها فقط, وبالتالي تقليل عدد التحاليل المطلوبة للمريض.

1. Introduction

The Data Mining (DM) approach considered as an important and widespread field, because it provides useful results in several areas. It is also easy to learn and provides many techniques that can be used in different ways. Therefore, it is considered one of the useful sciences in life from which humanity benefits every day in many discoveries. Thus, one of the most

*Email: sarasamir8880@yahoo.com

beneficiary fields is the medical field. Since, the more obtainable biological data, the more interest in bioinformatics to analyze for more different types of emerged data. First, bioinformatics was used to create and control databases for storing biological information. Then, with more data offered, the task of bioinformatics evolved to analyze them in order to have new hidden knowledge, including chemical tests, protein domains, protein structures and so on [1].

2. Related work

The scientists have many subjects for the DM techniques. According to [2], the diabetic patients' information from Ulster Community and Hospitals Trust (UCHT) from the year 2000 to 2004 is used to predict how well the patients' condition was controlled. The researchers used Feature Selection via Supervised Model Construction (FSSMC) to decide the important parameters in diabetic control, then the classification techniques, NB, IB1, and decision tree C4.5 were applied to the data. Then, in [3] the dataset is collected from the Ministry of Health, Saudi Arabia. Support vector machine algorithm was applied to investigate which case of treatment is efficient for each age category for diabetes patients. The researchers used the Oracle Data Miner tool to analyze the data. Thereafter, in [4], the researchers were used several algorithms (J48, basic logistics, and MLP) as machine learning approach to analyze real data from several Iraqi breast cancer cases in early detection hospitals using Weka data mining tool. And as a test choice, they employed 10-folds cross-validation, and a performance metric of a confusion matrix to evaluate the best of the suggested algorithms. The researchers also analyze if after several algorithm iterations, the error ratio decreases. It is lower in the case of MLP algorithm after 5-10 iterations rather than basic logistic, and J48 algorithms. On the other hand, in [5] the implementation of machine learning algorithms, where, the sample contains 370 employees in Iraq, it was preprocessed to represent the class attribute based on the gender value. Two DM approaches, the supervised attribute subset evaluator (CFS) with Greedy Stepwise as a search process, and Gain Ratio Attribute Evaluator with Ranker as a search method, they are utilized to pick the attribute for reducing the feature space. Also, the Apriori and association rule algorithms are then used to classify the key factors driving job apathy.

3. Methodology

In this article, the relationship among real biochemical tests has been analyzed to help in discovering how they affect each other. As well as, raw data have been used, which has been analyzed for the first time, also multiple preprocessing and DM algorithms for such type of dataset have been proposed.

3.1 Data Description

The investigated dataset was borrowed from a private Iraqi clinical biochemistry laboratory in Baghdad city, and recorded as a handwritten hard copy, then it had been converted to an electronic copy. The patients' cases had been described via 71 parameters. Whereas, the parameters had been classified as description into two groups. The first group consists of 66 parameters that present chemical tests, while the second group consists of 5 parameters that present personal information such as patient name, which already exists in the raw dataset. Also, the index, gender, age, and date have been added during preprocessing steps. Moreover, the number of patients' is 11000. Whereas, the number of females is 5343, while the number of males is 5657. Nevertheless, the number of adult is 10450, and the number of children is 550. The dataset is noisy (mixed data type, has huge missing values rate which equal to 94.8%). Also, non-labeled class, high dimension, and large variance in feature values. The tests details had been gained by interviewing a laboratory physician, also by the recorded documents and laboratory guidelines. The data are described as in Table 1. Accordingly, the Ch test is increasing in diabetes, chronic pancreatitis, and hypothyroidism, and it decreasing in chronic anemia, and malnutrition. While, Tri increases in the liver disease, and gout. Also, it decreases in case of malnutrition. As well as the HDL is called “beneficial” cholesterol because it keeps the arteries open and blood flows more smoothly. In other words, the higher the HDL level, the fewer incidents of arteriosclerosis, and angina. On the other hand, the LDL is called ” bad” cholesterol, since its increased amount in the blood causes accumulation of fatty deposits in the arteries, which leads to blood decreasing flow that could cause heart attack or stroke. Thus, LDL value could be calculated using equation 1 [7].

$$LDL = ch - HDL - \left(\frac{Tri}{5}\right) \qquad \dots\dots\dots\dots\dots\dots (1)$$

As well as, in the case of renal disease, Bu level is increasing in blood. In contrast, the level decreases in cases of liver disease due to its inability to form it [7]. However, Cr increases in some disease such as diabetes, high blood pressure. Nevertheless, there are two types of bilirubin test, Direct and indirect, are measured for an adult to help the doctor determine the treatment for liver disease. The amount of iron present in the blood varies during the day, so this leads to request other tests including Fe and TIBC. Also, TIBC high level means iron deficiency anemia which resulting from bleeding. The reasons for TIBC decline is the cancer of the digestive system. Further, coagulation profile contains PT which represents the clotting time, therefore this test is calculated and the doctor depends on its value in addition to the value of the INR test.

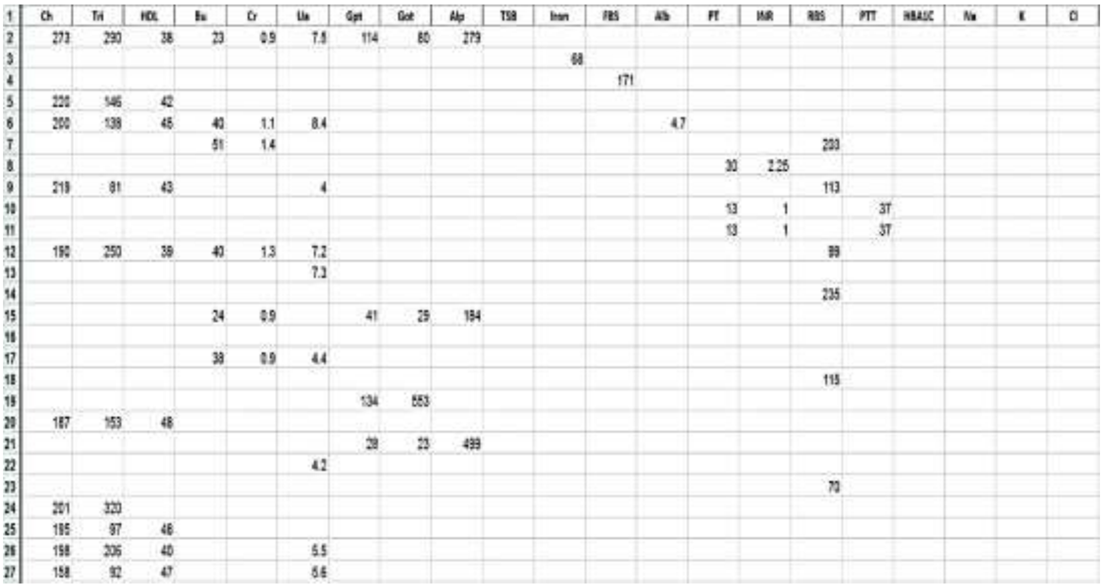
Table 1-Biomedical data description with normal values [6].

Field name	**Data type**	**Description**	**Normal value**	
Index	Int64	patient unique number	0-10999	
Name	Object	patient name		
Gender	Object	patient gender	M - F	
Age	Object	patient age	child - adult	
Date	Object	test date	DD-MM-YYYY	
Ch	Int64	Total Cholesterol, serum	<200 mg/dL	
Tri	Int64	Triglyceride, serum, mg/dL	Male 60 – 160	Female 40 – 140
HDL	Int64	high-density lipoprotein, mg/dL	Male 35 – 65	Female 35 – 70
LDL	float64	low-density lipoprotein	65 – 178 mg/dL	
Bu	Int64	Blood urea, plasma or serum	20 – 45 mg/dL	
Cr	float64	Creatinine, serum, mg/dL	Male 0.7 – 1.2	Female 0.5-1
Ua	float64	Uric acid, serum	3–7 mg/dL	
GPT	Int64	Glutamic pyruvic transaminase	5 – 65 U/L	
GOT	Int64	Glutamic oxaloacetic transaminase	<50 U/L	
ALP	Int64	Alkaline phosphatase, serum, U/L Age	Female	Male
	1 – 30 days		48 - 406	75 – 316
	1 month – 1 year		124 - 341	82 – 383
	1 – 3 years		108 - 317	104 – 345
	4 – 6 years		96 – 297	93 – 309
	7 – 9 years		69 – 325	86 – 315
	10 - 12 years		51 - 332	42 – 362
	13 – 15 years		50 - 162	74 – 390
	16 - 18 years		47 - 119	52 - 171
	20 – 50 years		42- 98	53- 128
	>Anni			

TSB	float64	Bilirubin, serum, mg/dL	Total 0.3 – 1.0	Direct 0.1 – 0.3	InDirect 0.2 – 0.7
Iron	Int64	Iron, serum	65 – 180 mg/dL		
FBS	Int64	Fast Blood Sugar	70 – 120 mg/dL		
ALB	float64	Albumin, serum	3.6 – 5.2 g/dL		
PT	float64	Prothrombin time	11 - 13.5 seconds		
INR	float64	international normalized ratio	0.8 – 1.1		
RBS	Int64	Random Blood Sugar	80 – 130 mg/dL		
PTT	Int64	Partial thromboplastin time	30 – 40 seconds		
HBA1C	float64	Hemoglobin A1C	4.2% – 6.2%		
Electrolytes, serum, mmol/L					
Na	Int64	Sodium, Natrium	136– 155		
K	float64	Potassium	3.6-5.5		
Cl	Int64	Chloride	98-111		
G6PD	Object	Glucose-6-phosphate dehydrogenase	Normal - deficient		
Urine stone analysis					
S-color	Object	color	brown, black, white, yellow		
S-size	float64	size, cm			
S-consistency	Object	consistency	solid		
S-ca	Int64	Calcium	found=1, not-found=0		
S-ox	Int64	Oxalate	found=1, not-found=0		
S-po4	Int64	Phosphate	found=1, not-found=0		
S-ua	Int64	Uric Acid	found=1, not-found=0		
S-carbonate	Int64	carbonate	found=1, not-found=0		
Amylase	Int64	Amylase, serum	<86 U/L		
Lipase	Int64	Lipase, serum	< 38 U/L		
Zinc	float64	Zinc, serum	72.6 – 127 mg/dL		
Ca	float64	Calcium, serum	8.5-10.5 mg/dL		
CRP	float64	C-reactive protein, serum	<5 mg/L		
Po4	float64	Phosphorous, serum, mg/dL	Children 4–7	Male 2.5–4.5	Female 1.5–6.8
24-hour urine test					
ur 24-cr	float64	Creatinine	1 – 2 g/24hr		
ur 24-ua	Int64	Uric acid	250 – 750 mg/24hr		
ur 24-ca	Int64	Calcium	100 – 350 mg/24hr		
ur 24-po4	float64	Phosphate	0.4 – 1.3 g/24hr		
ur 24-protein	float64	Protein	1.5 – 4.5 g/24hr		
ur 24-alb	float64	Albumin	<2.3 mmol/L		
ur 24-ox	float64	Oxalate	<0.50 mmol/L		
ur 24-citrate	Object	Citrate	positive – negative		
ur 24-cu	float64	Copper	20 – 50 mg/24hr		
ur 24 amylase	Int64	Amylase	1 - 17 U/24hr		
CK, CPK	Int64	Creatine kinase, serum, U/L	Children <225	Male <174	Female <140
TIBC	Int64	total iron-binding capacity, serum	250 – 400 mg/dL		

BJP	Object	Bence-Jonse protein	positive - negative	
LDH	Int64	lactate dehydrogenase, serum	207 – 414 U/L	
Mg	float64	Magnesium, serum, mmol/L	Children 1.7 – 2.3	Adult 1.6 – 3
GTT (1, 2, 3, 4, 5)	Int64	Glucose Tolerance Test (75gms)	70 – 120 mg/dL	
Complement components, serum, mg/dL				
C3 C4	Int64 float64	91 – 156 20 - 50		
Cu	float64	Copper, serum, mg/dL	Man (70-14) Women (80 – 155) Women Pregnancy (120 – 300) Children up to 1-year (80 – 190) Newborns (20 – 70)	
ACP	float64	Acid Phosphatase, serum	0.5–2.0 U/mL	
GGT, g-GT	Int64	Gamma-glutamyl transpeptidase, serum, U/L	Male <49	Female <32
TSP	float64	Total serum protein	6 – 8 g/dL	
Fe	Int64	Ferritin, serum, ng/mL	Male 30-400	Female 13-150
Lupus	Object	Lupus, Anti-lupus	Positive – negative	

The figure below show sample of the pure dataset for 27 patients, with the missing value and high dimensions:



1	Ch	Tri	HDL	Bu	Cr	Ua	Gpt	Got	Alp	TSB	Iron	FBS	Alb	PT	INR	RBS	PTT	HBA1C	Na	K	Cl
2	273	290	38	23	0.9	7.5	114	80	279												
3											68										
4												171									
5	220	146	42																		
6	200	138	46	40	1.1	8.4							4.7								
7				51	1.4											203					
8														30	2.25						
9	219	81	43			4										113					
10														13	1		37				
11														13	1		37				
12	190	250	39	40	1.3	7.2										99					
13						7.3															
14																235					
15				24	0.9		41	29	184												
16																					
17				38	0.9	4.4															
18																115					
19							134	553													
20	187	153	48																		
21							28	23	499												
22						4.2															
23																70					
24	201	320																			
25	195	97	46																		
26	198	206	40			5.5															
27	158	92	47			5.6															

Figure 1-Pure dataset excel sheet.

3.2 The Preprocessing Methods

In order to specify the preprocessing operations and DM algorithms, at the beginning it is important to understand the needs of the dataset. Python programming language was used to add the index feature, where a unique number was added for dataset indexing by Pandas library, to deal with dataset depending on the index (idx) feature rather than names., the age feature has been added to the dataset based on some tests by clinical laboratory physician support Moreover, as it has been assumed that the tests of children were (TSB, Ca, ALB, Bu, G6PD, RBS), while the remaining tests for adults. Also, the gender feature that was added depending on standard names in Iraq, assuming that common names are for females, because of their majority in Iraq. While, the date feature was added depending on the recorded date in the registry hard copy. Hence, null values should be removed. So, regarding the clinical laboratory dataset, the separating has been performed depending on similarities with features names (group of patients with the same tests), where a number of smaller separated datasets files has been created without null values. The resulting datasets is consisting of many outlier that have been removed by putting a thresholds, which is equal to 50 for sample size and 7 for number of features. Thereafter, the separate datasets have been grouped depending on several common features (biochemical tests) between them. The resulted groups of separate datasets were six. The first group has four common features (Ch, Tri, HDL, LDL) and 14 datasets. Then, the second group has one common feature (HBA1C) and 2 datasets. And, the third group has one dataset with (Iron, TIBC) features. While the fourth group has two common features (PT, INR) and 2 datasets. And, the fifth group has two common features (indirect, Direct) and 2 datasets. As well as, the sixth group has two common features (Bu, Cr) and 5 datasets. However, discretization process has been applied for each common feature to convert data type from numeric to nominal values relying on the standard reference of tests as explored in Table 2, then to be assumed as classes.

Table 2-Nominal values for assumed classes.

Tests	Nominal Values		
Ch	C1: if Ch ≤ 200 (Normal)	C2: if Ch > 200 (High)	
Tri	TM1: if Tri < 60 and Gender = “M” (Low)	TM2: if Tri ≥ 60 and Tri ≤ 160 and Gender = “M” (Normal)	TM3: if Tri > 160 and Gender =”M” (High)
	TF1: if Tri < 40 and Gender = “F” (Low)	TF2: if Tri ≥ 40 and Tri ≤ 140 and Gender = “F” (Normal)	TF3: if Tri > 140 and Gender = “F” (High)
HDL	HM1: if HDL < 35 and Gender = “M” (Low)	HM2: if HDL ≥ 35 and HDL ≤ 65 and Gender = “M” (Normal)	HM3: if HDL > 65 and Gender =”M” (High)
	HF1: if HDL < 35 and Gender = “F” (Low)	HF2: if HDL ≥ 35 and HDL ≤ 70 and Gender = “F” (Normal)	HF3: if HDL > 70 and Gender = “F” (High)
LDL	L1: if LDL < 65 (Low)	L2: if LDL ≥ 65 and LDL ≤ 178 (Normal)	L3: if LDL > 178 (High)
Cr	CM1: if Cr < 0.7 and Gender = “M” (Low)	CM2: if Cr ≥ 0.7 and Cr ≤ 1.2 and Gender = “M” (Normal)	CM3: if Cr > 1.2 and Gender =”M” (High)
	CF1: if Cr < 0.5 and Gender = “F” (Low)	CF2: if Cr ≥ 0.5 and Cr ≤ 1 and Gender = “F” (Normal)	CF3: if Cr > 1 and Gender = “F” (High)
Bu	B1: if Bu < 20 (Low)	B2: if Bu ≥ 20 and Bu ≤ 45 (Normal)	B3: if Bu > 45 (High)
Direct	D1: if Direct < 0.1 (Low)	D2: if Direct ≥ 0.1 and Direct ≤ 0.3 (Normal)	D3: if Direct > 0.3 (High)
indirect	in1: if indirect < 0.2 (Low)	in2: if indirect ≥ 0.2 and indirect ≤ 0.7 (Normal)	in3: if indirect > 0.7 (High)
INR	IN1: if INR < 0.8 (Low)	IN2: if INR ≥ 0.8 and INR ≤ 1.1 (Normal)	IN3: if INR > 1.1 (High)
PT	P1: if PT < 11 (Low)	P2: if PT ≥ 11 and PT ≤ 13.5 (Normal)	P3: if PT > 13.5 (High)
Iron	I1: if Iron < 65 (Low)	I2: if Iron ≥ 65 and Iron ≤ 180 (Normal)	I3: if Iron > 180 (High)
TiBc	TB1: if TiBc < 250 (Low)	TB2: if TiBc ≥ 250 and TiBc ≤ 400 (Normal)	TB3: if TiBc > 400 (High)
HBA1C	H1: if HBA1C < 4.2 (Low)	H2: if HBA1C ≥ 4.2 and HBA1C ≤ 6.2 (Normal)	H3: if HBA1C > 6.2 (High)

3.3 Proposed Feature Selection Methods

Features are mixture of noises and effective features. Therefore, feature selection technique was used to remove the noise, and as a result improving the accuracy by finding the highest impact features on the class value with less training time, and memory efficiency. Recursive Feature Elimination (RFE) which is a wrapper method, which starts with all dataset features, builds a model, and ignores the irrelevant feature according to the model. Then, a new model has been built using the rest features, and so on until a predetermined number of features are left, or reach high accuracy in case of using cross-validation [8]. Then, CART algorithm has been used as estimator for RFE, with criterion of 'gini', and random state =7. The 10-fold has been used to evaluate the model, so the accuracy for classes such as in Table 3. The RFE has been applied to the datasets which have at minimum two features in addition to the class, so the number of groups of datasets which, RFE has been applied to it, is three. Then, the assumed class with high feature selection accuracy has been selected to be the official class when apply CART algorithm.

Table 3-The results for RFE with Cross Validation as average of class's samples

Class name	**Ch**	Tri	HDL	LDL	**Bu**	Cr	**Direct**	indirect
10-fold accuracy	**0.94**	0.50	0.58	0.92	**0.88**	0.46	**0.95**	0.88

3.4 Classification and Regression Trees (CART) Algorithm

The Classification and Regression Trees (CART) approach constructs a binary tree, where each internal node denotes a condition on a feature, each of the two branches corresponds to a conditional outcome (true and false), and each leaf node denotes a class label. This algorithm chooses the "best" feature at each node to separate the data into individual classes depending on the criterion such as 'gini' [9]. Also, the 'gini' gain can be obtained by measuring the 'gini' index for all feature values of which belongs to the dataset. So, for dataset T, the 'gini' is determined as in equation 2 [10].

$$gini\ (T) = 1 - \sum_{j=1}^{n} p_j^2 \qquad \dots\dots\dots\dots\dots\dots..(2)$$

Where, n is the number of classes and p_j is the probability of different classes for the dataset samples. Gini split info, which measures the gini index for all feature values, which is determined according to equation 3:

$$gini_{split}\ (T) = \sum \frac{N_i}{N} gini(T_i) \qquad \dots\dots\dots\dots(3)$$

Where, i represents the $i-th$ feature value. And, the gain is the same, which is also called gini information gain (gini-gain).

3.5 Model Implementation and Evaluation

The CART algorithm has been implemented on the selected features datasets, in addition to the datasets with one feature and class. Then, hyperparameters tuning step which represent the search for a set of optimal hyperparameters, has been applied. The hyperparameters are set before training the model, such as splitting criterion, max-depth, and min-samples-leaf. Here grid search, which is comprehensive search within specified hyperparameters sets, has been used by determine sets of values for hyperparameters by define the range of possible values. Then define max-depth-set = (2,3,4) and min_sample_leaf_set = (0.05,0.06,0.07,0.08,0.09,0.1) such that hyperparameters space = {(2,0.05),(2,0.06),...,(4,0.1)} to choose the best hyperparameters values corresponding to larger cross validation accuracy. The grid search is suitable for low dimensions dataset because of time complexity [11]. 10-fold cross validation has been applied with the model, so the metrics for model evaluation for each class's samples such as (10-fold and testing accuracy, precision, recall, f1-score) are explained as Table 4. However, the bar plots for the features and classes relationships have been presented as Figures 1, 2, 3, 4, and 5, for discussion five patterns of the resulted relationships.

Table 4-The resulted measurements for CART model evaluation of classes samples.

Class name	Sample no.	10-fold accuracy	Testing accuracy	Class label	Precision	Recall	F1-Score
Ch	1	0.91	0.90	C1	0.94	0.89	0.92
				C2	0.85	0.92	0.88
	2	0.92	0.87	C1	0.86	0.96	0.90
				C2	0.91	0.73	0.81
	3	0.91	0.92	C1	1.00	0.89	0.94
				C2	0.80	1.00	0.89
	4	0.93	0.93	C1	0.88	1.00	0.93
				C2	1.00	0.86	0.92
	5	0.91	0.92	C1	1.00	0.86	0.93
				C2	0.85	1.00	0.92
Bu	1	0.90	0.90	B2	0.92	0.94	0.93
				B3	0.87	0.82	0.85
Direct	1	0.95	0.98	D2	0.96	1.00	0.98
				D3	1.00	0.97	0.98
INR	1	0.99	1.00	IN2	1.00	1.00	1.00
				IN3	1.00	1.00	1.00
	2	1.00	1.00	IN2	1.00	1.00	1.00
				IN3	1.00	1.00	1.00
Iron	1	0.94	1.00	I1	1.00	1.00	1.00
				I2	1.00	1.00	1.00
				I3	1.00	1.00	1.00

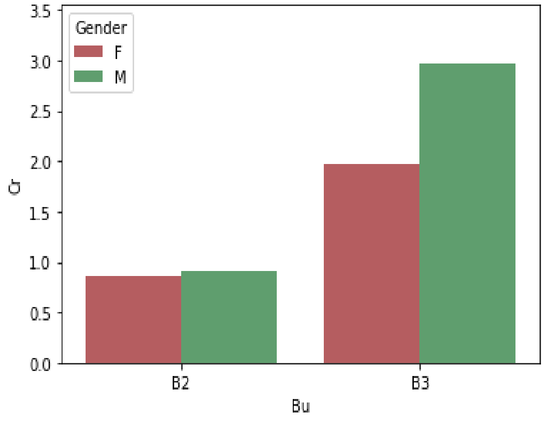


Figure 2-pattern of **Cr** feature with **Bu** class.

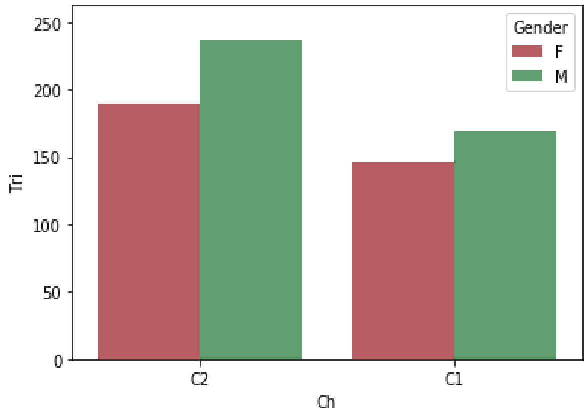
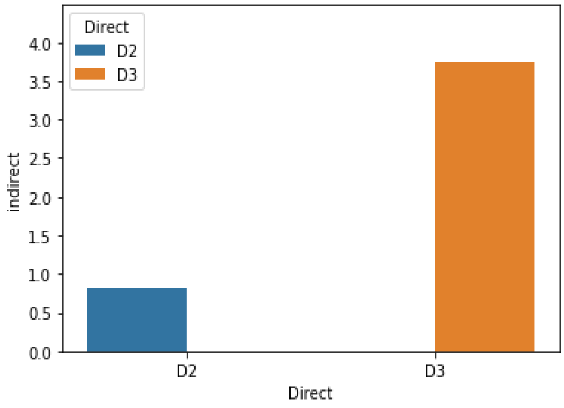


Figure-3 Pattern of **Tri** feature with **Ch** class.



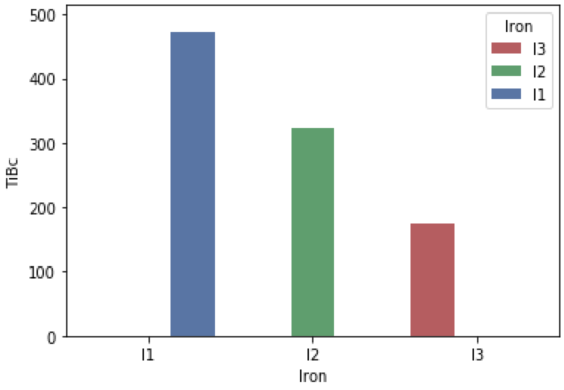


Figure 4-Pattern of **indirect** feature with **Direct** class. **Figure-5** Pattern of **TiBc** feature with **Iron** class.

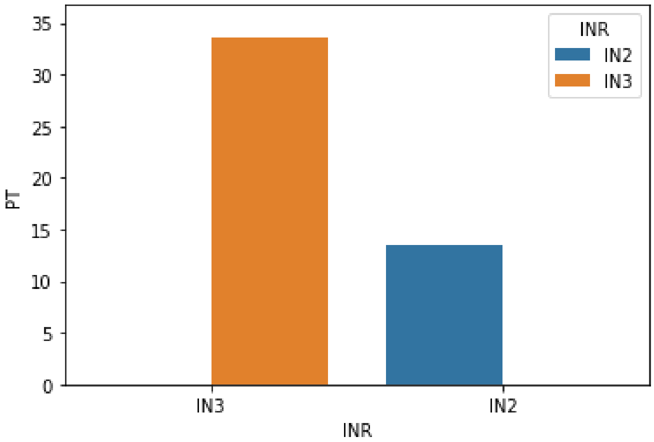
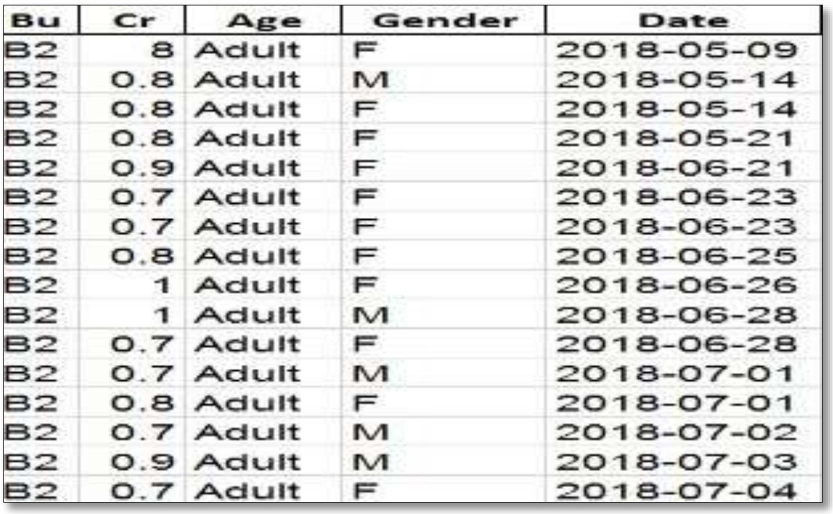


Figure 6-Pattern of **PT** feature with **INR** class.

The figure below show **Bu** class dataset sample after the previous processes:



Bu	Cr	Age	Gender	Date
B2	8	Adult	F	2018-05-09
B2	0.8	Adult	M	2018-05-14
B2	0.8	Adult	F	2018-05-14
B2	0.8	Adult	F	2018-05-21
B2	0.9	Adult	F	2018-06-21
B2	0.7	Adult	F	2018-06-23
B2	0.7	Adult	F	2018-06-23
B2	0.8	Adult	F	2018-06-25
B2	1	Adult	F	2018-06-26
B2	1	Adult	M	2018-06-28
B2	0.7	Adult	F	2018-06-28
B2	0.7	Adult	M	2018-07-01
B2	0.8	Adult	F	2018-07-01
B2	0.7	Adult	M	2018-07-02
B2	0.9	Adult	M	2018-07-03
B2	0.7	Adult	F	2018-07-04

Figure 7-Bu class sample excel sheet.

4. Results Discussion

The results for experiments explore that there is no relationship between total cholesterol test and Triglyceride tests with average of five results for 10-fold accuracy=**0.92** and testing accuracy=**0.91 (error=0.08, and 0.09)**. On the other hand, there is a positive relationship between Blood Urea test and Creatinine test with 10-fold accuracy=**0.90** and testing=**0.90 (error=0.1)**. Also, there is a positive relationship between international normalized ratio test and Prothrombin time test with 10-fold=**1.00** and testing=**1.00 (error=0)**. However, there is a positive relationship between Direct test and indirect test with 10-fold=**0.95** and testing=**0.98 (error=0.05, and 0.02**). Also, there exists an inverse relationship between Iron test and total iron-binding capacity test discovered with 10-fold=**0.94** and testing=**1.00 (error=0.06, and 0)**. Furthermore, the Gender has effect in determine the normal values from the standards of the biochemical tests. Meanwhile, the age parameter was the same for all patterns and is similar for adult, with considering that the resulting patterns were for adults only. However, there was no specific date for these patterns.

5. Conclusions

This paper presented the experiments that could be applied upon this type of dataset. For purpose of discovering hidden relationships patterns among biochemical tests, and detecting what are the helpful algorithms and what are not. Patterns that have been discovered could be useful in diagnostic issues without need for more tests. This would support Iraqi medical physicians in decision making process. Additionally, the proposed algorithms will help the researchers in manipulating such type of data that was not analyzed previously. Consequently, the Classification and Regression Trees (**CART**) algorithm has been noticed as useful in the clinical field. Indeed, the preprocessing phase had been regarded as a very important part for this kind of dataset investigation due to its owing high noise, null values, and high complex raw data.

References

[1] F. Martin-Sanchez and K. Verspoor, "Big data in medicine is driving big changes," *Yearbook of medical informatics*, vol. 9, no. 1, pp. 14, 2014. [Online]. https://doi.org/10.15265/IY-2014-0020

[2] Y. Huang, P. McCullagh, N. Black, and R. Harper, "Feature selection and classification model construction on type 2 diabetic patients' data," *Artificial intelligence in medicine*, vol. 41, no. 3, pp. 251–262, 2007. [Online]. https://doi.org/10.1016/j.artmed.2007.07.002

[3] A. A. Aljumah, M. G. Ahamad, and M. K. Siddiqui, "Application of data mining: Diabetes health care in young and old patients," *Journal of King Saud University-Computer and Information Sciences*, vol. 25, no. 2, pp. 127–136, 2013. [Online]. https://doi.org/10.1016/j.jksuci.2012.10.003

[4] S. F. Behadili, M. S. Abd, I. K. Mohammed, and M. M. Al-SAYYID, "Breast cancer decisive parameters for Iraqi women via data mining techniques," *Journal of Contemporary Medical Sciences*, vol. 5, no. 2, 2019.

[5] M. S. Abd and S. F. Behadili, "Recognizing job apathy patterns of Iraqi higher education employees using data mining techniques," *Journal of Southwest Jiaotong University*, vol. 54, no. 4, 2019. [Online]. https://doi.org/10.35741/issn.0258-2724.54.4.27

[6] S. Fuggle, "Clinical biochemistry reference ranges hand-book,"*National Health Service, Leeds, UK*, 2018.

[7] M. Crook, *Clinical biochemistry and metabolic medicine*. CRC Press, 2013. [Online]. https://doi.org/10.1201/b13295

[8] A. C. M¨uller, S. Guido *et al*., "*Introduction to machine learning with Python: a guide for data scientists*," O'Reilly Media, Inc.", 2016. [Online]. https://www.oreilly.com/library/view/introduction-to-machine/9781449369880/

[9] J. Han, J. Pei, and M. Kamber, "*Data mining: concepts and techniques,"* Elsevier, 2011. [Online]. https://doi.org/10.1016/C2009-0-61819-5

[10] M. Li, "Application of cart decision tree combined with pca algorithm in intrusion detection," in *2017 8th IEEE International Conference on Software Engineering and Service Science (ICSESS).* IEEE, 2017, pp. 38–41. [Online]. https://doi.org/10.1109/ICSESS.2017.8342859

[11] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikitlearn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.