DOI: http://dx.doi.org/10.21123/bsj.2022.19.2.0385

Data Mining Techniques for Iraqi Biochemical Dataset Analysis

*Sarah Sameer** *Suhad Faisal Behadili*

Computer Science Department, College of Science, University of Baghdad, Baghdad, Iraq.
Corresponding author: sarahsameer@scbaghdad.edu.iq, suhad.f@sc.uobaghdad.edu.iq
ORCID ID: https://orcid.org/0000-0002-6897-2864, https://orcid.org/0000-0003-4617-8464 +9647702583537

Received 13/7/2020, Accepted 19/1/2021, Published Online First 20/9/2021, Published 1/4/2022

Abstract:

This research aims to analyze and simulate biochemical real test data for uncovering the relationships among the tests, and how each of them impacts others. The data were acquired from Iraqi private biochemical laboratory. However, these data have many dimensions with a high rate of null values, and big patient numbers. Then, several experiments have been applied on these data beginning with unsupervised techniques such as hierarchical clustering, and k-means, but the results were not clear. Then the preprocessing step performed, to make the dataset analyzable by supervised techniques such as Linear Discriminant Analysis (LDA), Classification And Regression Tree (CART), Logistic Regression (LR), K-Nearest Neighbor (K-NN), Naïve Bays (NB), and Support Vector Machine (SVM) techniques. CART gives clear results with high accuracy between the six supervised algorithms. It is worth noting that the preprocessing steps take remarkable efforts to handle this type of data, since its pure data set has so many null values of a ratio 94.8%, then it becomes 0% after achieving the preprocessing steps. Then, in order to apply CART algorithm, several determined tests were assumed as classes. The decision to select the tests which had been assumed as classes were depending on their acquired accuracy. Consequently, enabling the physicians to trace and connect the tests result with each other, which extends its impact on patients' health.

Keywords: Biomedical, Classification And Regression Tree (CART), Data mining, Hierarchical clustering, K-means.

Introduction:

The biochemical tests make a wide correlation with human organs functions, as well as the imbalanced glands and hormones. Therefore, they help in discovering the diseases and future risks. The used tests types are include, blood tests that are related to chronic diseases and include lipid, liver function, renal function, carbohydrates, bone markers, electrolytes, anemia, pancreatic function, coagulation, and cardiac function tests. Likewise, the previous researches that have used biochemical tests have diagnosed breast cancer, diabetes, and heart diseases for several specific tests such as in [1-3]. While in this paper, the dataset have more biochemical tests over the previous works, also there is no specified disease diagnosed. Data mining (DM) as an important science in the biomedical analysis domain, it participates to discover hidden knowledge and rules from the dataset by extracting patterns, clusters, or relationships [4-5]. The big challenge in this search is the null values ratio, and how to cover this issue by proposing algorithm to solve it. Also, the assumption classes were necessary for work with supervised algorithms, then classify data in order to analysis tests to discover patterns of relationships. The remainder of this paper is organized as follows, section 2 presents related work, section 3 presents methodology, section 4 presents data mining techniques, 5 presents CART algorithm, 6 present model implementation, and lastly discussions and conclusions.

Related Work

There are many researches addressed various issues of DM approaches. Hence, in [1] they apply machine learning algorithms (J48, simple logistic, and Multiplayer Perceptions (MLP)) using Weka DM tools, they observe real data that are acquired from several Iraqi hospitals of early detection for breast cancer. The researchers also applied $10-folds\ cross-validation$ as a test option and a confusion matrix as a performance metric to choose the optimal one from proposed algorithms. Also, the error ratio was tested, which decreased after 5-10 iterations of algorithm execution in the case of MLP algorithm rather than simple logistic, and J48 algorithms. While in [2], the Pima Indian Diabetes dataset was used to improve the system accuracy for classifying diabetes disease. The researchers proposed a hybrid of machine learning algorithms, Self-Organizing Map (SOM), Principal Component Analysis (PCA), and neural network (NN) for clustering, noise removal, and classification tasks respectively. Whereas in [3], The diabetic patients' information from Ulster Community and Hospitals Trust (UCHT) for the years 2000 to 2004, which are used to predict how well the patients' condition was controlled. The researchers used feature selection via supervised model construction (FSSMC) and optimization of ReliefF to decide the important parameters in diabetic control, then the classification techniques Naive Bayes (NB), IB1, and decision tree C4.5 were applied to the data. In [4] they used two benchmark medical datasets from the UCML repository to find risk patterns. They proposed Mining Optimal Risk pattern sets (MORE) algorithm for risk patterns extraction. So, they compared the proposed method with the decision tree C5.0 algorithm, a commercial version of C4.5, and variant association rule mining-based approaches (Apriori). However, in [6] the dataset is collected from the Ministry of Health, Saudi Arabia. It supports vector machine algorithms that were applied to investigate which treatment case is more efficient for each age category of diabetes patients. The researchers used the Oracle Data Miner tool to analyze their data. As well as, in [7] the used data of Multiparameter Intelligent Monitoring Intensive Care II (MIMIC-II) physiology patients are investigated via DM techniques, logistic regression, NN, DT, and K-Nearest Neighbor for predicting death cases within the next 24 hours. So, NN and logistic regression generated better results, while the configuring parameters contributed to model success. In turn, the researchers of [8] have been experimenting disease prediction tools from eight different predictive classifiers. These classifiers are NB, decision tree J48, Instance-Based Learning (IBK), Sequential Minimum Optimization (SMO), MLP, DT Reduced Error Pruning (REP) Tree, Projective Adaptive Resonance Theory (PART), and Random Forest (RF). They used data for breast cancer, diabetes, heart disease, Tuberculosis (TB), and liver disorder available from UCI, and another source of data that contains Human Immunodeficiency Virus (HIV). These data sets are gathered from Amana hospital in Dares-salaam, Tanzania. They developed two-hybrid systems, the first one used Wisconsin Breast Cancer dataset with combinations of SMO+RF+IBK, and SMO+RF+MLP, which introduced good performance. Whereas, in the second one they used an HIV dataset with a combination of SMO + J48 + MLP that exhibited good performance. Indeed, the $10-fold\ cross-validation$ was used as a test option, while the confusion matrix as a performance metric. Furthermore, in [9] the used dataset is acquired via a general hygiene questionnaire form, which was designed and distributed for 200 students of two high schools in Baghdad city. The questions reflect general environmental health characteristics. The proposed approach consists of three DM techniques, Apriori, association rule mining, and NB respectively. They encoded and analyzed the data using the Weka DM tool for uncovering hidden relationships among their parameters.

This research use real mixture data, from private laboratory, which consist of many blood tests that no predecessor research has analyzed. While the past researches usually used part of those tests to diagnostic specific disease. Noticed from the previous works, that supervised techniques useful for such type of data.

Material and Methods:

Material:

In this study, the relationship between real biochemical tests have been analyzed to help in discovering how they affect each other, in order to help in reduce the tests number, as well as the tests cost. Raw data have been used, there are no research have used same of our dataset, also we have proposed powerful preprocessing and DM algorithms for such type of dataset. The investigated data have been manipulated through many experiments to discover the relationship between the tests, and the affected patterns on the class value. The shape of the data has clearly affected the steps of preprocessing type. In the next section, some issues will be addressed like the data description, experiments that have been applied, and the results of each experiment.

Data Description

The investigated dataset was borrowed from a private Iraqi 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 by 71 parameters. Whereas, the parameters had been classified 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. However, the index, gender, age, and date have been added during preprocessing steps. Moreover, the patients' number is 5609. However, the females' number is 2691, while the males' number is 2918. Nevertheless, the adult number is 5461, and the children number is 148. The tests details had been gained by interviewing a laboratory physician, also by the recorded documents and laboratory guidelines. The data could be described as in Table 1.

Table1. Biomedical data description with normal values.

Field name	Data type	description	normal value		
Index Name Gender Age Date	Int64 object object object object	patient unique number patient name patient gender patient age test date	 M - F child - adult 		
Ch	Int64	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/Ur	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 ALT, SGPT AST, SGOT	float64 Int64 Int64	Uric acid, serum Aminotransferase, serum alanine Aminotransferase, serum aspartate	3–7 mg/dL 5 – 65 U/L <50 U/L		
ALP	Int64	alkaline phosphatase, serum, U/L	Age	Female	Male
			1 - 30 days	48 - 406	75 – 31
			1mon - 1year	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 FBS ALB PT INR RBS PTT HBA1C	Int64 Int64 float64 float64 float64 Int64 Int64 float64	Iron, serum Fast Blood Sugar Albumin, serum prothrombin time international normalized ratio Random Blood Sugar Partial thromboplastin time Hemoglobin A1C	65 – 180 mg/dL 70 – 120 mg/dL 3.6 – 5.2 g/dL 11 - 13.5 seconds 0.8 – 1.1 80 – 130 mg/dL 30 – 40 seconds 4.2% – 6.2%		
		Electrolytes, serum, mmol/L			
Na K Cl	Int64 float64 Int64	Sodium, Natrium Potassium Chloride	136– 155 3.6-5.5 98-111		
G6PD	object	Glucose-6-phosphate dehydrogenase	normal - deficient		
		Stone analysis			
S-color S-size S-consistency	object float64 object	color size, cm consistency			

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
S.Amylase	Int64	Amylase, serum	<86 U/L
S.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	Int64		91 – 156
C4	float64		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

Often, there are no specific symptoms of high cholesterol. The extra cholesterol may be stored in the arteries as plaques, which leads to narrow arteries. Therefore, patients have to check the other cholesterol tests that include Tri, HDL, and LDL. Where, Tri is another type of cholesterol, which its level increasing leads to arteries hardening. As well as, LDL is called "bad" cholesterol, since its excessed amount in the blood leads to heart attack or stroke. Thus, LDL value could be calculated using equation 1.

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

Meanwhile, the Bu/Ur test is used to diagnose kidney diseases. In the case of kidney disease, the Bu/Ur level will be high in the blood. In contrast, the Bu/Ur decreases in liver disease cases due to its inability to form it. Whereas, any disease that causes kidney weakness could lead to high Cr test, such as diabetes, high blood pressure. Likewise, the TSB test as a total is for children to measure the percentage of bilirubin. When TSB increasing, mean there is jaundice. Despite the fact that TSB test is calculated to direct/indirect for an adult to measure the liver enzymes, and for supporting the doctors in determining the liver disease treatment. Meanwhile, the prothrombin protein produced by the liver, is one of many blood factors that helps to clot appropriately. Where, the PT test determines how quickly blood clots. It's often performed along with PTT test, which looks at another set of factors. When patients take blood thinners such as heparin, warfarin, and aspirin, then the prothrombin time test results will be expressed as INR test [10].

Methods of Work:

There are several DM techniques implemented, the unsupervised techniques were used firstly because the investigated laboratory dataset has no specific class. However, the applied experiments on this dataset are discussed in the below subsections.

Experiment 1

In Experiment 1 the clustering through hierarchical technique was applied by orange platform V3.22.0. Thereafter, the hierarchical technique clustered patients into subgroups, these subgroups are then merged into larger groups, then formatting the hierarchical tree. The distance computation have default imputation for null values, either by the average of row or of column. Hence, agglomerative strategies had been used with parameters, Euclidian distance metric, normalization, no pruning, whereas the selection was manual, and the linkage was average based as formulated in equation 2 [11]:

$$mean_i \,\|a_i \;-\; b_i\,\| \ldots\ldots\ldots\ldots\ldots\ldots.(2)$$

Where, a_i and b_i are all objects in clusters a and b. Anyway, this technique offers a dendrogram that is a tree like-diagram, which records the sequences of merges. The drawback is the determination of the clusters number should be manually, here k has been set as 3. Therefore the silhouette algorithm for detection the cluster number were used. The produced result as represented in Fig. 1.

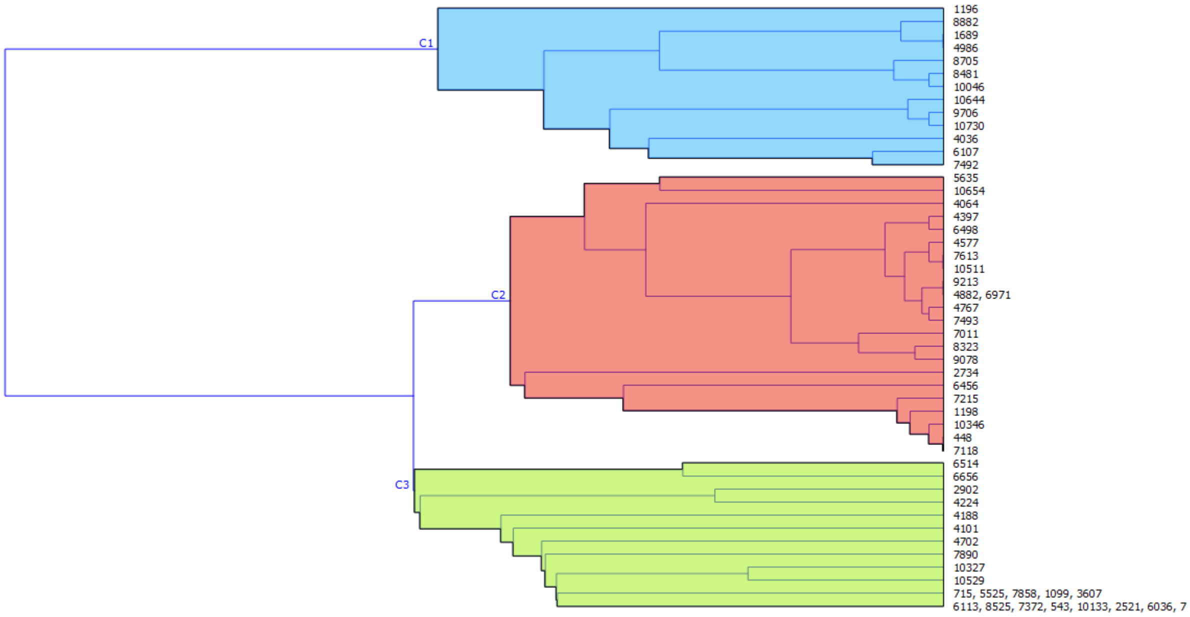
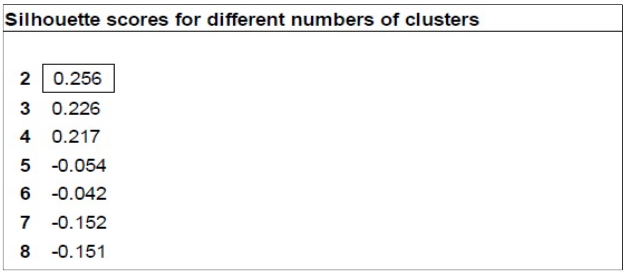


Figure 1. Hierarchical clustering algorithm using Orange platform.

Experiment 2

In experiment 2 a Silhouette algorithm was applied to determine clusters number for *k-means* approach. The used platform was orange too. Thus, Silhouette has a drawback because it cannot be applied to null values. As a consequence, filling was based on the average of each feature. Also, this algorithm doesn't work with more than 5000 instances. Hence, the result explored two clusters with a high score. Then, the *k-means* algorithm was applied with k=2. The first produced cluster contained 2873 patients, while the second cluster contained 2127 patients. Accordingly, the silhouette algorithm scores are shown in Fig. 2.



Silhouette scores for different numbers of clusters

2	0.256
3	0.226
4	0.217
5	-0.054
6	-0.042
7	-0.152
8	-0.151

Figure 2. The Silhouette algorithm produced two clusters of the highest score.

However, for a dataset D of n objects, if D is split to k clusters C_1, ..., C_K. Then, for each object $o \in D$, $a(o)$ is calculated as the average distance between

o and all other objects in the cluster that o belongs to. On the other hand, $b(o)$is the minimum average distance from o to all clusters that o do not belong to. Then, for $o \in C_i$ $(1 \leq i \leq k)$. In this manner, the distances are determined as shown in equations 3 and 4 respectively [12].

$$a(o) = \frac{\sum_{o` \in C_{i,o\neq o`}} \quad dist(o,o`)}{|C_i|-1} \ldots\ldots\ldots\ldots.(3)$$

$$b(o) = \left\{\frac{\sum_{o` \in C_j} \quad dist(o,o`)}{|C_j|}\right\} \ldots\ldots\ldots.(4)$$

Then, the Silhouette coefficient of o is determined as in equation 5 [12].

$$s(o) = \frac{b(o) - a(o)}{\{a(o), b(o)\}} \ldots\ldots\ldots\ldots..(5)$$

Where, the result of equation 5 is ranging between -1 and 1. While, the value of $a(o)$ indicates the cluster agglomerating to which object belongs. Whenever this value becomes smaller, hence the cluster is more agglomerated. Meanwhile, the value of $b(o)$indicates the object far away degree from other clusters. Whenever the value becomes larger, then the object will be more separated from other clusters [12.]

Experiment 3

The dataset had many null values, which cannot be filled by approximate values, because this will lead to loss information and give wrong result. Also, the dataset features cannot be reduced by any dimensional reduction technique, because it may lead to probable loss important features. Since, CART will be used, the dataset did not need normalization process. Thus, a supervised principle has been used by applying multi-steps beginning with preprocessing which is the most important step, and then several DM algorithms have been examined on the dataset in order to select the highest accuracy one. In this way, these steps are described in the following sections:

Preprocessing Step

In order to perform the preprocessing phase, python v. 3.7.0. have been used. Thus, adding the index feature was the first step, where a unique number was added for dataset indexing. So, an index object has been used, where pandas library v. 0.25.1 provides it. Moreover, the age feature has been added to the dataset based on some tests by clinical laboratory physician support, as it has been assumed that the tests of children were (TSB, Ca, ALB, Bu, G6PD, RBS), while the remaining tests for adults. However, Adding Age Feature has been determined as in algorithm 1.

Algorithm 1: Adding Age Feature

```
Input: X
Output: X with age feature
Function CreTestLst(X):
Set L1←X[TSB].notna(), X[Ca].notna(), X[ALB].notna(), X[Bu].notna(), X[G6PD].notna(), X[RBS].notna()
Set L2←X not in L1
Set N← len(X)
ComTestAdd(X, L1,L2,N)
return X
End Function
Function ComAgeAdd(X,L1,L2,N):
Set X[age]←nan
for i←0 to N do
    if X in L1 then
        Set X[age]←child
    end
    if X in L2 and not in L1 then
        Set X[age]←adult
    end
end
End Function
```

Whereas X is the indexed dataset, N is the length of X, $L1$ is the list for children tests names, $L2$ is the list for adult tests names, and i is a temporary variable. 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, as determined in algorithm 2.

Algorithm 2: Adding Gender Feature

```
Input: X
Output: X with gender feature
Function CreNamesLst(X):
Set L1←[standard names for male]
Set N← len(X)
ComGenAdd(X, L1,N)
return X
End Function
Function ComGenAdd(X, L1,N):
Set X[gender]←nan
for i←0 to N do
    if X in L1 then
        Set X[gender]←'M'
    else Set X[gender]←'F'
    end
end
End Function
```

Whereas, *X* is the indexed dataset with age feature, *N* is the length of *X*, *L1* is the list of standard male names, and *i* is a temporary variable. While, the date feature was added depending on the recorded date in the registry hard copy. This feature was added as in algorithm 3.

Algorithm 3: Adding Date Feature

```
Input: X,FDate
Output: X with date feature
Function ComDateAdd(X,FDate,N,j,k):
Set X[date]←nan
Set N←len(X)
for i←0 to N do
if i > = j and i < = k  then
Set X[date]←FDate
else
Set  j←k+1
Set k←k+100
Set FDate← New Date
end
end
return X
End Function
```

Where, *X* is the indexed dataset with age and gender features, *FDate* is the first date in the registry hard copy, *New Date* is the next date, *N* is the length of *X*, and *i,j,k* are temporary variables. Hence, data cleaning is necessary because null values make a big problem in the data analysis field. Therefore, they should be removed. Regarding the clinical laboratory dataset, the splitting has been performed depending on similarities with features names, where a number of smaller separated datasets files has been created without null values. The number of resulting datasets is 609, which consist of many outliers datasets that have been removed by putting a thresholds, which is equal to 50 for sample size and 3 for number of tests features at least. This procedure has been performed via algorithm 4.

Algorithm 4: Splitting Dataset

```
Input: X
Output: Split datasets as excel files
Function CreColLst(X):
Set N←len(X)
Set j←0
Set m←0
for i←0 to N do
Set k[j]← X[i].columns names
Set j←j+1
if k[j] not in x and len(k) > 2 then
Set x[m]← k[j]
Set m←m+1
end
end
return x
End Function
Function CreSplit(X):
Set m←0
for j← 0 to 608
for i←x[0] to x[608] do
X[i].drop row that contains a null value at column names
Set K[m]←X[i].index values
Set m←m+1
X[i].write to excelfiles (j)
X.drop(K)
end
end
return excelfiles
End Function
Function DelFiles(excelfiles):
for i←0 to 608 do
Set a←len(excelfiles(i))
if a < 50 then
Delete(excelfiles(i))
end
end
End Function
```

Whereas *X* is the indexed dataset with age, gender, and date features, *N* is the *X* length, *k* is the columns names list, *x* is the list of column names list without repeating, and *K* is the index list, which is same column names list in *X*, and *i, j, m, a* are temporary variables. As well as, data visualization step was needed because the human grasp a lot of information from diagrammatic representation. Thus, it is better to visualize dataset splitting through visualization techniques. So, a scatter plot function in the Seaborn library v. 0.10.1 has been used for plotting splitted datasets. Thereby, it assists in clarifying the relation between features, such as represented in Fig. 3. Thereafter, the splitted datasets have been grouped depending on several common features between them. Where the chemical tests should be splitted from personal information. The result was six groups of splitted datasets, the first group has four common features (Ch, Tri, HDL, LDL). Then, the second group has three common features (FBS, RBS, HBA1C). And, the third group has two common features (Iron, TIBC). While the fourth group has three common features (PT, INR, PTT). And, the fifth group has two common features (TSB, Direct). As well as, the sixth group has two common features (Bu, Cr). Accordingly, the grouping process was designed as in algorithm 5.

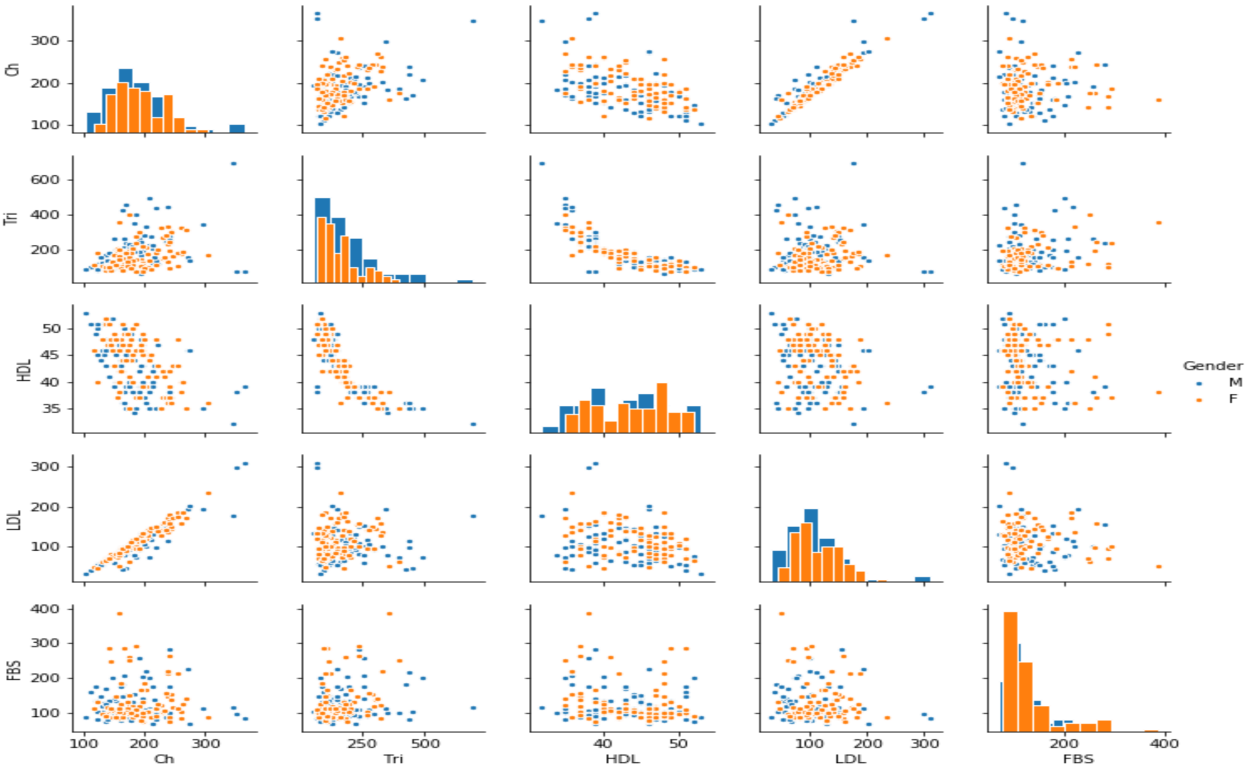


Figure 3. Plot of one splitted data sets with four common features (Ch, Tri, HDL, LDL,FBS)

Algorithm 5: Grouping Dataset

```
Input: F
Output: list of common features
Function CalCF(F):
    for k ←1 to len(F) do
        Set f ← F[k]
        ComFea(f)
    end
End Function
Function ComFea(f):
    Set X←read max length dataset from f
    Set L←X.column names
    L.delete elements from 0 to 5
    For i←0 to len(f)
        Set X1←read dataset from f
        Set L1←X1.columns names
        L1.delete elements from 0 to 5
        for j←0 to len(L)
            If L[j] not in L1 then
                L.remove(j)
            end
        end
    end
    return L
End Function
```

Whereas, *F* is the groups file, *f* is one of the group elements, *X* is the longest *excel file* in *f*. While *X1* is a temporary variable for the rest *excel files*, *L* and *L*1 are the lists of columns names, and *k*,*i*,*j* are temporary variables. And then, data discretization process has been applied for each common feature to convert data type from numeric to nominal values relying on the normal reference of tests as explored in Table 2, then to be assumed as classes. The discretization technique provided by python was unhelpful, thus algorithm 6 has been developed for this process using Table 2.

Table 2. nominal values for assumed classes

TESTS	NOMINAL VALUES		
Ch	C1: if Ch < = 200 (Normal)	C2: if Ch > 200 (High)	
Cr	Cr1: if Cr < 0.7 (Low)	Cr2: if Cr ≥ 0.7 and Cr ≤ 1.2 (Normal)	Cr3: if Cr >1.2 (High)
Direct	D1: if Direct < 0.1 (Low)	D2: if Direct ≥ 0.1 and Direct ≤ 0.3 (Normal)	D3: if Direct > 0.3 (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)
TSB	T1: if TSB < 0.2 (Low)	T2: if TSB ≥ 0.2 and TSB ≤ 0.7 (Normal)	T3: if TSB > 0.7 (High)

Algorithm 6: Discretization Common Features

```
Input: $F$,$L$,$NR1$,$NR2$,$nominal1$,$nominal2$,$nominal3$
Output: $X$ with nominal values at common features
Function CalDisc($F$,$L$, $NR1$,$NR2$,$nominal1$,$nominal2$,$nominal3$):
    for $k \leftarrow 1$ to len($F$) do
        Set $f \leftarrow F[k]$
        Disc($f$,$L$ $NR1$,$NR2$,$nominal1$,$nominal2$,$nominal3$)
    end
End Function
Function Disc($f$,$L$ $NR1$,$NR2$,$nominal1$,$nominal2$,$nominal3$):
    Set $X \leftarrow$read dataset from $f$
    Set $L1 \leftarrow X$.columns names
    if $L1 == L$ then
        For $i \leftarrow 0$ to len($X$)
            if $X[L1,i] < NR1$ then
                $X[L1,i]$.replace('$nominal1$')
            else if $X[L1,i] \geq NR1$ and $\leq NR2$ then
                $X[L1,i]$.replace('$nominal2$')
            else $X[L1,i]$. replace('$nominal3$')
            end
        end
    end
End Function
```

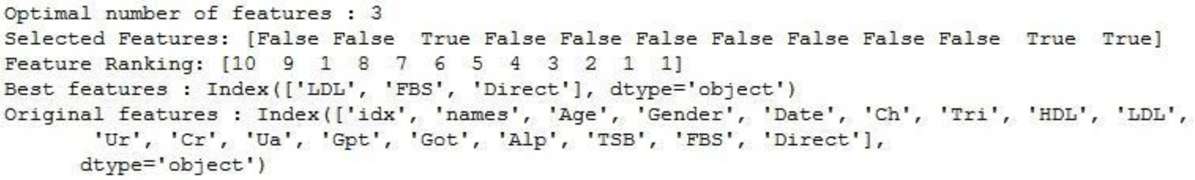
Where, F is the groups file , L is the common features list , $NR1$ and $NR2$ are normal range values for common features (tests), $nominal1$,$nominal2$, and $nominal3$ represent nominal values for assumed classes (features), f is one of the group elements, X is *excel file* in f, $L1$ is the list of features names, and k,i are temporary variables.

Feature Selection

The feature selection technique is very useful in improving the accuracy by finding the highest impact features on the class value with less training time, and memory efficiency. However, feature selection algorithms are subdivided into supervised, and unsupervised. Hence, the two feature selection techniques which provided by sklearn library v. 0.22.1 have been used on a splitted dataset sample. Firstly variance-threshold technique has been used, which is unsupervised that employs the reducing feature principle, where predetermined probability is used in calculating threshold according to equation 6 [13], which is used at the same time in calculating feature variance when using feature probability:

$$Var[X] = p(1-p) \dots\dots\dots\dots..(6)$$

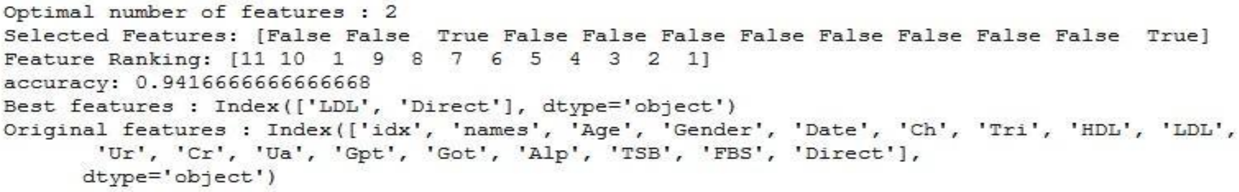
Where, p reflects predetermined probability, and feature values probability. If feature variance values less than a threshold, then it should be ignored. It was unhelpful in the clinical laboratory dataset, because the features have high variance, they are numeric with different values. Eventually, Recursive Feature Elimination (RFE) which is a supervised technique that starts with all dataset features, builds a model, and ignores the irrelevant features 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 [14]. Thus, it needs to determine the features number and the model. Then, three used for features number, and the decision tree classifier with a criterion of 'gini' as a model. It was applied to one of the split datasets that assumed (Ch) feature as a class. So, the result supports the selected features list, features ranking list. Whereas, the chosen features have rank=1, and best features list, such as represented in Fig. 4.



```
Optimal number of features : 3
Selected Features: [False False  True False False False False False False False  True  True]
Feature Ranking: [10  9  1  8  7  6  5  4  3  2  1  1]
Best features : Index(['LDL', 'FBS', 'Direct'], dtype='object')
Original features : Index(['idx', 'names', 'Age', 'Gender', 'Date', 'Ch', 'Tri', 'HDL', 'LDL',
       'Ur', 'Cr', 'Ua', 'Gpt', 'Got', 'Alp', 'TSB', 'FBS', 'Direct'],
      dtype='object')
```

Figure 4. The RFE results for one of the splitted datasets.

In RFE with $cross - validation$ (RFECV) of *10-fold* , tuning the predetermined number of features will be automatic. Thus, an accuracy metric results such as in Fig. 5.



```
Optimal number of features : 2
Selected Features: [False False  True False False False False False False False False  True]
Feature Ranking: [11 10  1  9  8  7  6  5  4  3  2  1]
accuracy: 0.9416666666666668
Best features : Index(['LDL', 'Direct'], dtype='object')
Original features : Index(['idx', 'names', 'Age', 'Gender', 'Date', 'Ch', 'Tri', 'HDL', 'LDL',
       'Ur', 'Cr', 'Ua', 'Gpt', 'Got', 'Alp', 'TSB', 'FBS', 'Direct'],
      dtype='object')
```

Figure 5. RFECV technique results for one of the splitted datasets.

Data Mining Algorithms

Several DM algorithms that Sklearn library provided have been used firstly. The $10-fold\,cross-validation$ has been used as a performance metric to evaluate each algorithm. Also the dataset has been splitted to 80% for training and 20% for testing evaluation. Actually, LR, LDA, CART, K-NN, NB, and SVM have been applied on the splitted dataset, where (Ch) feature as the assumed class with 61 sample. Hence, the results were the accuracy mean of *10-fold*, and the testing accuracy as shown below in Table 3. Indeed it's obvious that CART algorithm has the highest accuracy of 0.93, and 0.92 respectively. Thus it has been used as an analysis model.

Table 3. Accuracy results for DM algorithm.

DM algorithms	LR	CART	NB	LDA	SVM	KNN
10-fold accuracy	0.91	0.93	0.85	0.90	0.59	0.89
Testing accuracy	0.77	0.92	0.69	0.85	0.69	0.77

The concept of cross-validation mean split the dataset into k parts of equal size then, k-1 using for training, and the rest 1 for testing. The process of cross-validation repeat k times in order to avoid bias in result. Commonly in evaluation of algorithms, researchers either use the k-fold or training-testing accuracy, which mean split the dataset for two parts, training and testing. Generally, the size of training be 75% or 80%, so for testing using 25% or 20% of the dataset. This process run only one time.

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. The tree becomes binary depending on the feature selection measure. Some feature selection measures, such as the ‘gini’ index, enforce the resulting tree to be binary. Others, like information gain, which allow multiway splits [12,14]. Also, the ‘gini’ feature gain can be obtained by measuring the ‘gini’ index for all feature values of which belongs to the dataset. As well as, when the pruning is not used, then the building process of decision tree will select the gini-gain of the smallest node, which is the branching point until the sub-datasets belong to the same class or all the features are used in building the tree. So, for dataset T, the ‘gini’ is determined as in equation 7 [14]:

$$gini\,(T) = 1 - \sum_{j=1}^{n} \quad p_j^2 \dots\dots\dots\dots\dots\dots..(7)$$

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 8 [14]:

$$gini_{split}\,(T) = \sum \quad \frac{N_i}{N} gini(T_i) \dots\dots\dots\dots(8)$$

Where i represents the $i-th$ feature value. And, the gain is the same, which is also called gini information gain (gini-gain). Likewise, for CART, i = (1,2), get gini gain in binary split according to equation 9 [14]:

$$gini_{split}\,(T) = \frac{N1}{N} gini(T1) + \frac{N2}{N} gini(T2) \dots\dots..(9)$$

because the algorithm relies on the principle of the binary tree, so concerning continuous data, a discretization process used on the data by considering it not continuous for one sample, such that if there are N samples, this means that there is $N-1$ of the split results, the right sub-tree represents values bigger than, while the left sub-tree represents values less and equal to the parent node. In order to reduce the computations for the discretization process, the feature have been arranged in ascending order and select midpoint as a division point which divides the data into two parts, likewise calculate the gini-gain for each possible division point. But here the improvement with this algorithm is to calculate gini-gain for only the distinct values of the classification attribute change. Then choose the value with the lowest gini-gain as the best separation point.

Model Implementation

In this paper, the model has been built using the CART algorithm with $k-fold\,cross-validation$ performance metric. The parameters for CART were ‘gini’ for criterion, 7 for random_state value, and 3 for max_depth value. *K-fold* algorithm parameters were 10 for K and 7 for random_state. Hence, Fig. 6 below shows the model design.

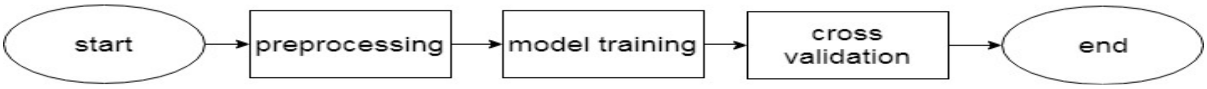


Figure 6. CART model design flow chart.

However, each group of common features has been expanded to a number of groups equal to a number of common features. Then each common feature in the group has been assumed as a class. After applying the CART algorithm in the orange platform and python language, the same result has been obtained. But, the accuracy of some assumed classes was low, therefore they have been ignored. Then, high accuracy assumed classes were, Ch, Cr, TSB, PT, INR, and Direct. The tree has been visualized by using the matplotlib v 3.1.1 library as shown in the following Fig. 7-13.

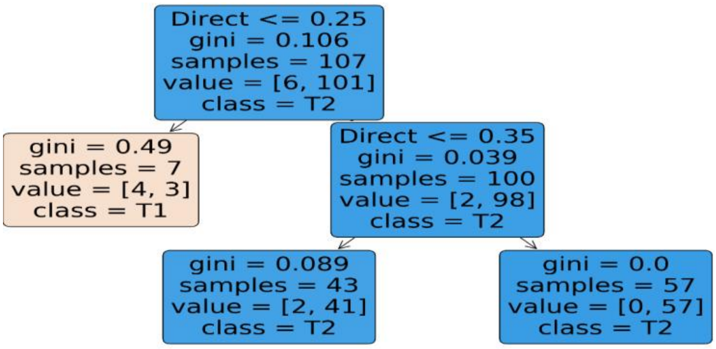


Figure 7: pattern of Direct feature with TSB class.

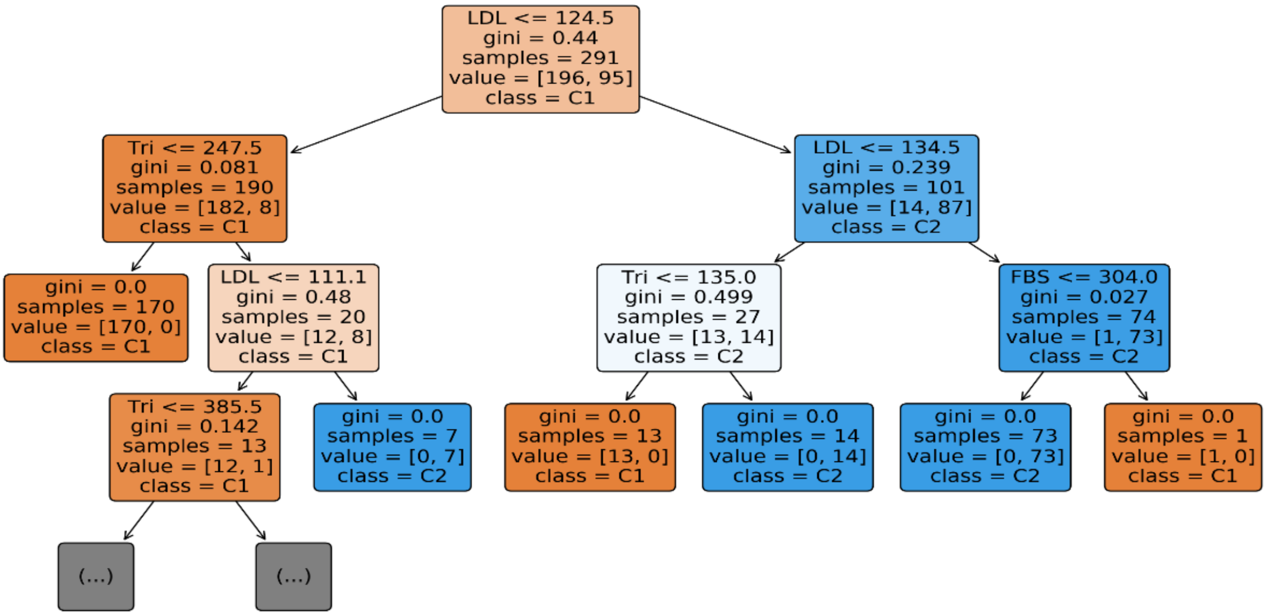


Figure 8: Pattern of LDL and Tri features with Ch class.

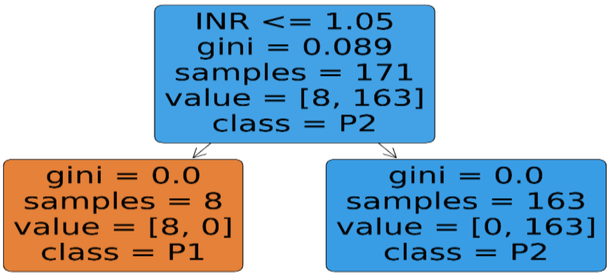


Figure 9: Pattern of PT feature with INR class.

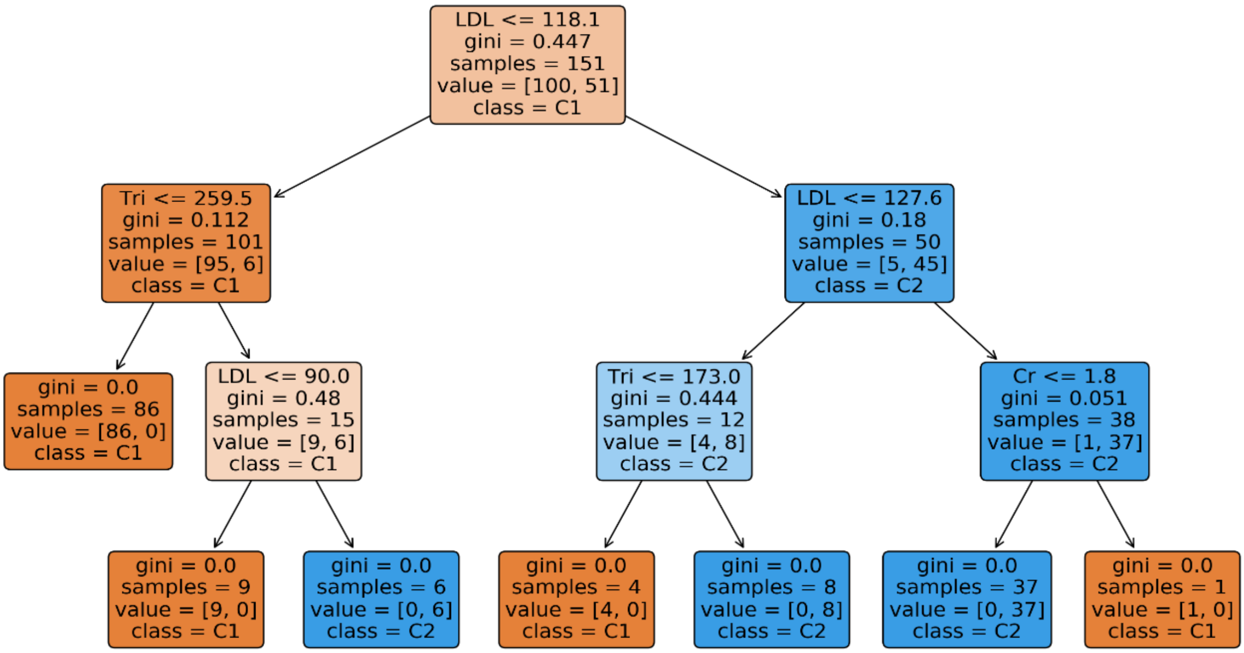


Figure 10. Pattern of LDL, Tri, and Cr features with Ch class.

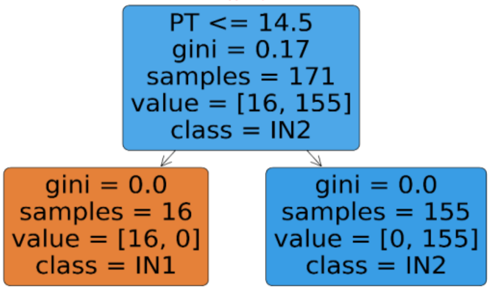


Figure 11. Pattern of INR feature with PT class.

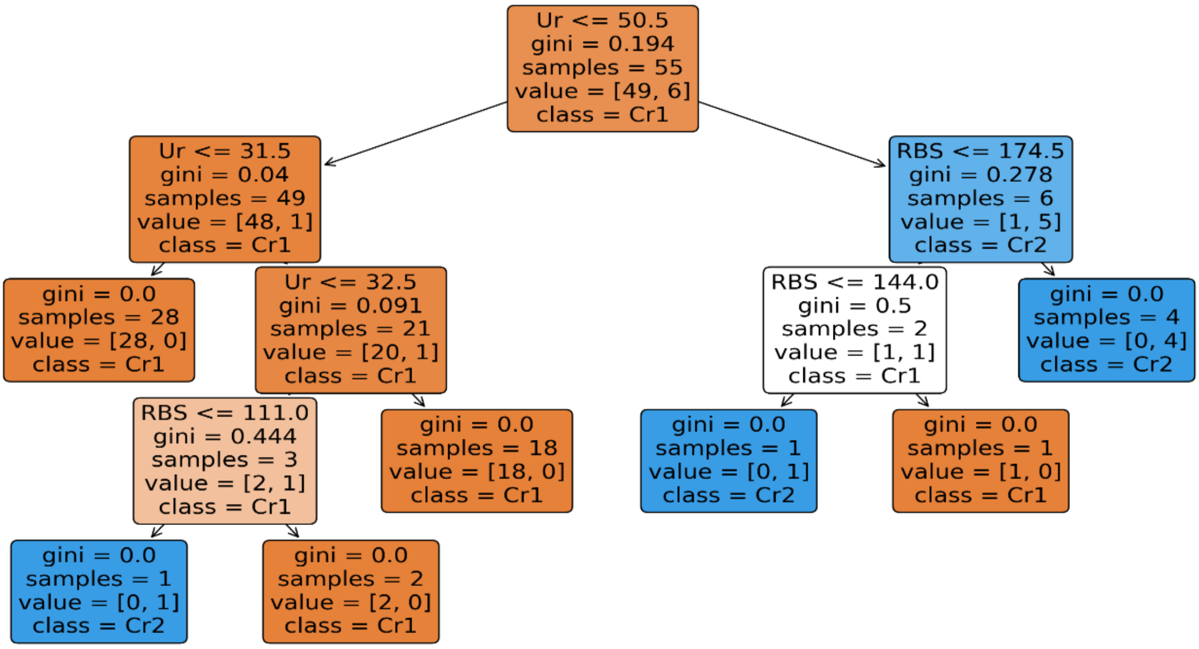


Figure 12. Pattern of Ur feature with Cr class.

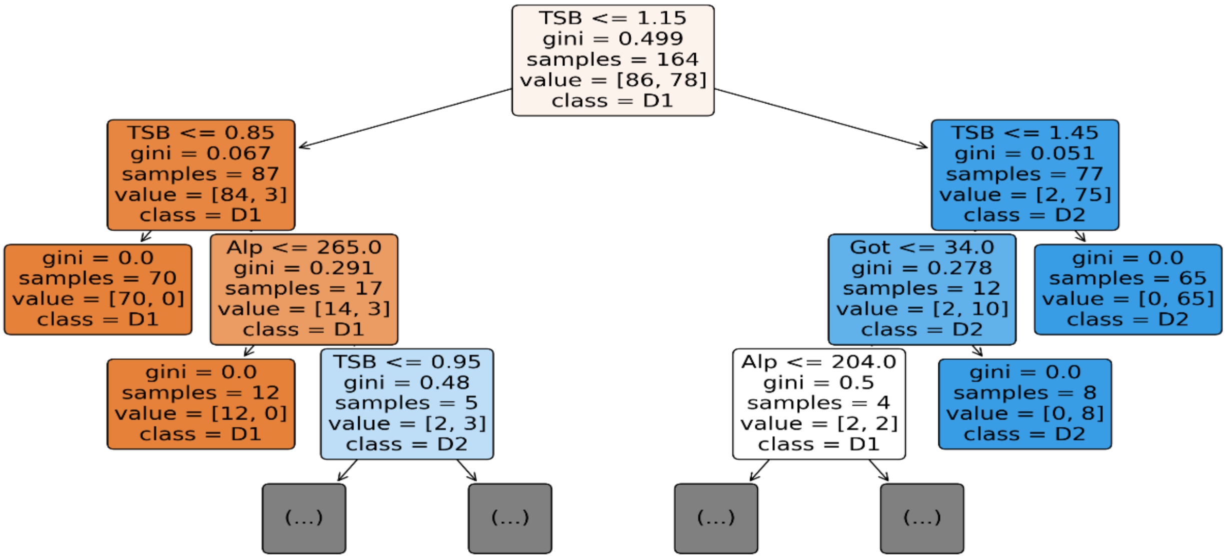


Figure 13. Pattern of TSB feature with Direct class.

Results and Discussion:

The resulting accuracy of patterns was determined as $10-fold\ cross-validation$. The resulting patterns were LDL value decreasing does not affect Ch value, it stays normal, while LDL increasing leads to Ch increasing with an accuracy of 0.97 and 0.96. Also, Tri decreasing does not affect the value of the Ch with an accuracy of 0.97. While, the decrease of Cr value affects by increasing the Ch value with an accuracy of 0.96. Although, the Bu/Ur value does not affect the Cr value (there is no relation between them) with an accuracy of 0.93. Furthermore, the increasing of Direct does not affect the TSB value with an accuracy of 0.95. On the other hand, the decrease in TSB value affects Direct by decreasing it, while the increase in TSB does not affect direct value with an accuracy of 0.97. The INR value increasing does not affect the PT value, while when decreasing leads to low PT with an accuracy of 1.0. Also, when PT value decreasing leads to low INR, and when increasing, the value of the INR stays normal with an accuracy of 1.0 [10].

Conclusions:

This paper presented the experiments which could be applied to this type of dataset for discovering the patterns of relationships between biochemical tests, and detecting what are the helpful algorithms and what are not. The patterns that have discovered could be helped in diagnostic problems without need to more tests to help the Iraqi medical physician to take decisions. While, the proposed algorithms will help the researchers in such type of data that did not analyzed previously, because it acquired from the private Iraqi laboratory. Finally, it could be said that the theoretical concept for such type of study contribute in the future for disease diagnostics. The Classification and Regression Trees (CART) algorithm has been noticed as useful in the clinical field, according to its high gained accuracy with seed=7, and tree pruning, reversely, the SVM failed in analysis. Also, the preprocessing phase is a very important part of this kind of dataset investigation due to its high noise, null values, and high complex raw data. The discovered patterns may be helpful in detecting any health case without requiring more tests, since disease patterns could be discovered as future work by physician help. The suggestions for more research on similar dataset types as studying environmental pollution, and how they affect human health. But more details needed, including determining the place of residence to know where pollution lies according to regions in addition to the place of birth. For example, Lung cancer, heart disease and other diseases related to environmental pollution, according to the type of pollution.

Authors' declaration:

- Conflicts of Interest: None.
- We hereby confirm that all the Figures and Tables in the manuscript are mine ours. Besides, the Figures and images, which are not mine ours, have been given the permission for re-publication attached with the manuscript.
- Ethical Clearance: The project was approved by the local ethical committee in University of Baghdad.

Authors' contributions statement:

Sarah Sameer Rasheed is a master student in the Department of Computer Science in the College of Science at the University of Baghdad,

Baghdad/Iraq. She has a B.Sc. from the University of Baghdad, Baghdad/Iraq.
She is employee in the ministry of the labor and social affairs.
Suhad Faisal Behadili is a professor in the Department of Computer Science in the College of Science at the University of Baghdad, Baghdad/Iraq. She has a Ph.D. from the LITIS at Normandie University - Le Havre/ France. She currently editorial member of the American Journal of Information Science and Technology. She is a program committee member in several international conferences, and a reviewer for IJS and IASET journals. She has published numerous technical papers, undergraduate/Postgraduate teaching, and outreach.

References:

1. Behadili SF, Abd MS, Mohammed IK, Al-SAYYID MM. Breast cancer decisive parameters for Iraqi women via data mining techniques. JOCMS. 2019 Apr 19;5(2).
2. Nilashi M, Ibrahim O, Dalvi M, Ahmadi H, Shahmoradi L. Accuracy improvement for diabetes disease classification: a case on a public medical dataset. Fuzzy Inf. Eng. 2017 Sep 1;9(3):345-57. DOI: https://doi.org/10.1016/j.fiae.2017.09.006
3. Huang Y, McCullagh P, Black N, Harper R. Feature selection and classification model construction on type 2 diabetic patients' data. Artif Intell Med. 2007 Nov 1;41(3):251-62. DOI: 10.1016/j.artmed.2007.07.002
4. Li J, Fu AW, Fahey P. Efficient discovery of risk patterns in medical data. Artif Intell Med. 2009 Jan 1;45(1):77-89. DOI: 10.1136/svn-2017-000101
5. Wasan SK, Bhatnagar V, Kaur H. The impact of data mining techniques on medical diagnostics. Data Sci. J. 2006;5:119-26. DOI: http://doi.org/10.2481/dsj.5.119
6. Aljumah AA, Ahamad MG, Siddiqui MK. Application of data mining: Diabetes health care in young and old patients. JKSUCI. 2013 Jul1;25(2): 127-36. https://doi.org/10.1016/j.jksuci.2012.10.003
7. Salcedo-Bernal A, Villamil-Giraldo MP, Moreno-Barbosa AD. Clinical data analysis: An opportunity to compare machine learning methods. Procedia Comput Sci. 2016 Jan 1;100(100):731-8. DOI: 10.1016/j.procs.2016.09.218
8. Diwani SA, Yonah ZO. A novel holistic disease prediction tool using best fit data mining techniques. IJCDS. 2017 Mar 1;6(02):63-72. DOI: http://dx.doi.org/10.12785/IJCDS/060202
9. Mustafa TK, Abd MS. Proposed approach for analysing general hygiene information using various data mining algorithms. IJS. 2017;58(1B):337-44.
10. Crook M. Clinical biochemistry and metabolic medicine. 8th ed. London. CRC Press, 2012. DOI https://doi.org/10.1201/b13295
11. Drab K, Daszykowski M. Clustering in analytical chemistry. J AOAC Int. 2014 Jan 1;97(1):29-38. DOI:https://doi.org/10.5740/jaoacint.SGEDrab
12. Han J, Kamber M, Pei J. Data mining concepts and techniques. 3rd ed. Elsevier; 2011 Jun 9.
13. Müller AC, Guido S. Introduction to machine learning with Python: a guide for data scientists. " O'Reilly Media, Inc."; 2016 Sep 26.
14. Li M. Application of CART decision tree combined with PCA algorithm in intrusion detection. In2017 8th IEEE International Conference on Software Engineering and Service Science (ICSESS) 2017 Nov 24 (pp. 38-41). IEEE. DOI:10.1109/ICSESS.2017.8342859

تقنيات تنقيب البيانات لتحليل مجموعة البيانات البيوكيميائية العراقية

سهاد فيصل البهادلي **سارة سمير**

قسم علوم الحاسبات، كلية العلوم، جامعة بغداد، بغداد، العراق.

الخلاصة:

يهدف هذا البحث إلى تحليل ومحاكاة بيانات تحاليل الكيمياء الحيوية الحقيقية للكشف عن العلاقات فيما بين التحاليل ، وكيف يؤثر كل منها على الآخرين. تم الحصول على البيانات من مختبر الكيمياء الحيوية العراقي الخاص. كذلك فإن هذه البيانات لها أبعاد عديدة ذات معدل مرتفع من القيم الخالية وأعداد كبيرة من المرضى. بعد ذلك ، تم تطبيق العديد من التجارب على هذه البيانات بدءًا بتقنيات غير خاضعة للرقابة مثل التجمعات الهيكلية وك-الوسائل ، ولكن النتائج لم تكن واضحة. ثم تم تنفيذ خطوة المعالجة المسبقة ، لجعل مجموعة البيانات قابلة للتحليل من خلال تقنيات خاضعة للإشراف مثل التحليل التمييزي الخطي **(LDA)** ، وشجرة التصنيف والانحدار **(CART)** ، والانحدار اللوجستي **(LR)** ، و ك-اقرب جار **(K-NN)** ، و نايف بايز (**NB)** ، وتقنيات آلة ناقل الدعم **(SVM)**. يعطي **CART** نتائج واضحة بدقة عالية بين الخوارزميات الستة الخاضعة للإشراف. من الجدير بالذكر أن خطوات المعالجة المسبقة تتطلب جهودًا ملحوظة للتعامل مع هذا النوع من البيانات ، نظرًا لأن مجموعة البيانات الخالصة بها العديد من القيم الصفرية بنسبة 94.8٪ ، ثم تصبح 0٪ بعد تحقيق خطوات المعالجة المسبقة. ثم ، من أجل تطبيق خوارزمية **CART** ، تم افتراض العديد من الاختبارات المحددة كفئات. قرار اختيار الاختبارات التي تم افتراضها على أنها فئات كانت تعتمد على دقتها المكتسبة. وبالتالي ، تمكين الأطباء من تتبع وربط نتائج الاختبارات مع بعضها البعض ، مما يوسع تأثيرها على صحة المرضى.

الكلمات المفتاحية: الطب الحيوي، شجرة التصنيف والانحدار (CART)، استخراج البيانات، التجمعات الهرمية، ك-الوسائل.