Hybrid Feature Selection Algorithm to Support Health Data Warehousing

Md. Badiuzzaman Biplob [1,2]; Shahidul Islam Khan [1,3]; Galib Ahasan Sheraji [1]; Jubayed Ahmed Shuvo [1]

[1]Department of CSE, International Islamic University Chittagong
[2]Department of CSE, Chittagong University of Engineering & Technology
[3]Department of CSE, Bangladesh University of Engineering & Technology

biplob.cse45@gmail.com; nayeemkh@gmail.com;
galibahasan@yahoo.com; jubayedsr@gmail.com

Abstract. Large volumes of data are being generated each day in healthcare. In addition,these huge amounts of data from healthcare datasets cause the issue of proper knowledge discovery. Currently, data integration is an approach, which is increasingly utilized by healthcare data specialists for analyzing the information and data mining. "Which features or attributes should we use to integrate data for data warehouses"-is a difficult question to answer. It requires deep knowledge of the problem domain. Automatic feature selection is the process of selecting a subset of relevant features automatically for later use. In this paper, we proposed a method using four random forest based feature selection algorithm and domain knowledge. Experimental results show that our hybrid method can select a required number of features from a large set of attributes.

Keywords: Hybrid Feature Selection, Feature Selection, Data Integration, Data Warehouse

1 Introduction

As of late, data becomes bigger and speedier each day which makes it challenging for data researchers to separate and decipher the unpredictability of data in order to discover new information. In addition, a similar issue is also looked by health data system whether its data contain a large amount of health knowledge [1]. This is extremely helpful for the symptomatic motivation to help medicinal services establishments in explaining different social insurance inquire about issues [2].

In most healthcare data mining, the best way to enhance the data quality is by ignoring a sector, which may give better output close to our expectations. Rather, the majority of them just focused on choosing reasonable learning calculations. For example, Social insurance dataset contains many features that can affect the expected execution. In a similar way, the issue of the large data includes if its being found in high spatial of learning and great process differing nature [3, 5]. Feature selection, as projected by Singh et. al, implies the strategy together with the collection of set of features from particular attributes[5]. In any case, the most insightful models are keeping up a strategic way from the reasonable procedures for choosing the best feature. Accordingly, the probability thickness limit of the component vector space is lacking in the midst of the grouping task [6].

Understanding the standing of creating the class arrangements of data, the objective of this examination is to suggest a system based on the hybrid feature decision methods that can redesign the desired model of patients' data for grabbing a superior characterization result.

2 Related Works

Healthcare data Mining (HDM) is a technique for separating useful data and models from a large number of information [2]. This will be engaged around the techniques of delivering quality data. The technique for making efficient data is basically declaring the learning quality that can impact the arrangement procedure [7]. Data quality is frequently delivered in the midst of the preparing ready stage in record mining strategy. Hence, the part-alternative method is regularly used to require care of high dimensional data issue.

Feature selection includes four straightforward steps that set age, set analysis, stopping standard and result approval **[4]**.Initially, in the set age, the new applicant of feature subsets is conveyed using given procedure**.** Around then, the competitor subsets assessed and differentiated and past subsets in light of assessment rule**.** The past grasp sets will be eliminated if the last made subset is ideal**.** The two methodologies reiterated to the reason once a stopping model is fulfilled**.** Ultimately, the endorsement procedure on genuine information is foreseen to approve the best**-**chose grasp subsets**.** The Plain Steps of feature Assortment process (see Fig. 6)

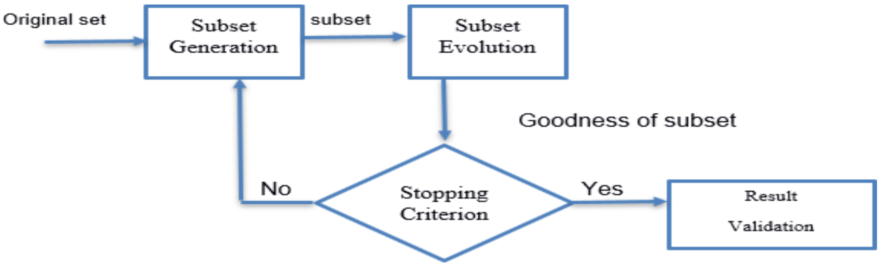


Fig.1. the Simple Steps of Feature Selection Procedure

3 PROPOSED FRAMEWORK

As to contemplate on feature selection that has been analyzed previously, the arranged system on smearing to illustrate feature selection method to clarify the high spatial data in social welfare is shown. [10] The system begins with data pre-planning as showed up in (Fig. 2).

The agenda of data pre-processing is to change the rough information into gainful data. In the middle of this strategy, the segment determination systems were associated with an upgrade to the idea of data. The four figuring used for Consolidating it will deal with the issue of high spatial data and give imposing accuracy occurs [5]. The methodology of the suggested structure is cleared up in the consequent area. The procedure of feature selection Technique appears in Fig. 2. Our Proposed hybrid feature selection algorithm is presented below:

Input: N features

Output: M features,

Where N>M
Steps:

1. Input N features
2. Using Feature selection with correlation and random forest classification find out the effective feature by domain knowledge
3. Using Univariate feature selection and random forest classification find out the effective feature by each attribute scores.
4. Using Recursive feature elimination (RFE) with random forest finds out the effective feature by running the output at 100times.
5. Using Recursive feature elimination with cross-validation and random forest classification find out the effective feature by running the output at 100 times.
6. Collect the common features from each Algorithm
7. Get M output, Where N>M

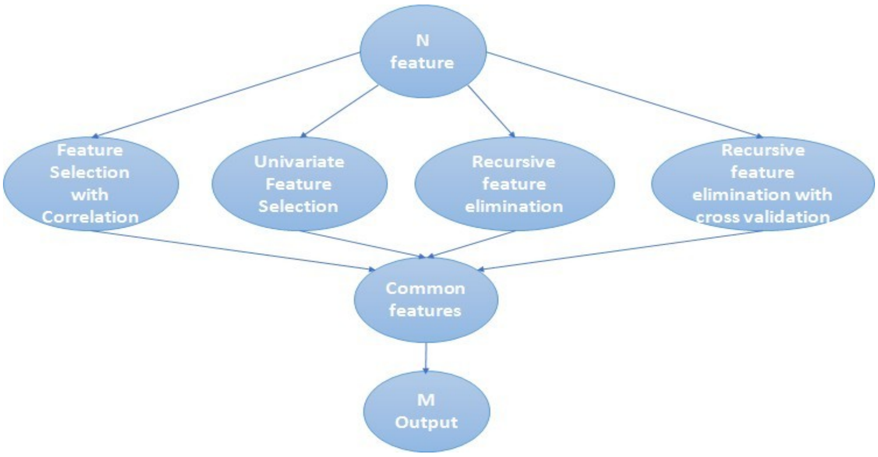
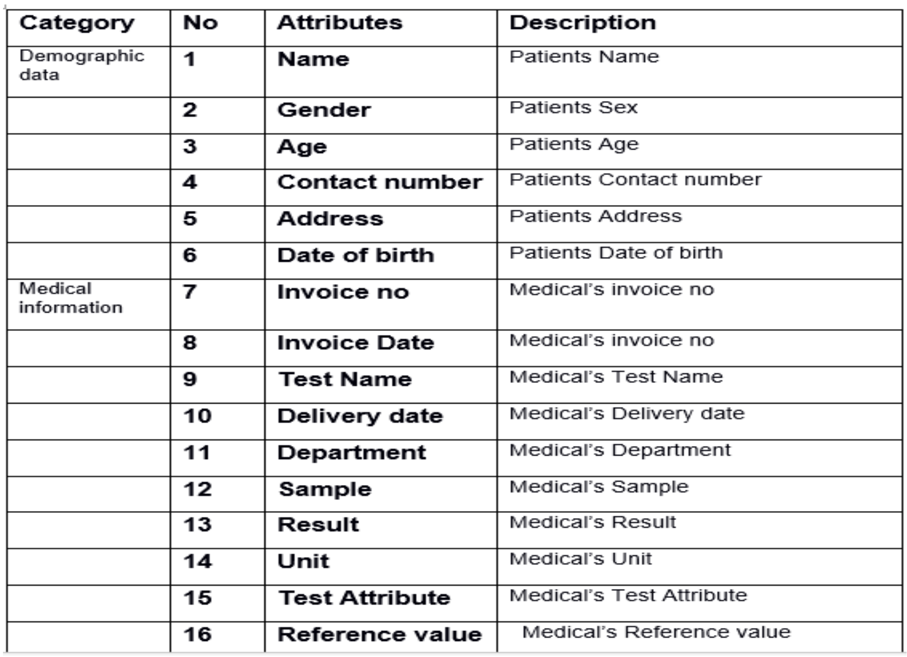

N feature
Feature Selection with Correlation
Univariate Feature Selection
Recursive feature elimination
Recursive feature elimination with cross validation
Common features
M Output

Fig.2. Proposed Framework “Feature Selection Technique

4 Description on Healthcare Dataset and Algorithm

There are several types of healthcare dataset and also many attributes in the dataset. It is needed to find out which features or attributes should we use to integrate data for data warehouses. For this purpose, we are proposing an algorithm where we are showing (see Fig.2) how to effectively find out the main feature from the dataset. In our dataset 16 types of attributes (see Table. I). Before implementing our algorithm we convert our dataset Categorical Values to numerical Values to easily find out the main feature from the dataset. In the first algorithm, find out the effective feature by domain knowledge. In the second algorithm, Univariate feature selection and random forest classification find out the effective feature by each attribute scores. The first and second algorithm output is not randomly generated. For this reason, we run the code only one time. However, in the third and fourth algorithm because of their random output, we run this code 100 times to find out the effective feature. Finally, showing the most common features with numbering from these four algorithms.

Table I. Description on healthcare dataset



Category	No	Attributes	Description
Demographic data	1	**Name**	Patients Name
	2	**Gender**	Patients Sex
	3	**Age**	Patients Age
	4	**Contact number**	Patients Contact number
	5	**Address**	Patients Address
	6	**Date of birth**	Patients Date of birth
Medical information	7	**Invoice no**	Medical's invoice no
	8	**Invoice Date**	Medical's invoice no
	9	**Test Name**	Medical's Test Name
	10	**Delivery date**	Medical's Delivery date
	11	**Department**	Medical's Department
	12	**Sample**	Medical's Sample
	13	**Result**	Medical's Result
	14	**Unit**	Medical's Unit
	15	**Test Attribute**	Medical's Test Attribute
	16	**Reference value**	Medical's Reference value

1. Feature selection with correlation and random forest classification
   - Select the feature at correlation based and used random forest supervised classification algorithm.
2. Univariate feature selection and random forest classification
   - The single selection of features works by choosing the best features on the basis of uniform statistic testing. It can be considered an estimator preprocessing step. Scikit-learn exposes routines for feature selection as objects implementing the transform method.
   - SelectKBest [11] eliminates everything except the K maximum counting features.
   - The SelectKBest class just scores the features utilizing a capacity.
   - SelectKBest [11] select the best K includes that have the extreme importance with the objective variable. It takes two parameters as

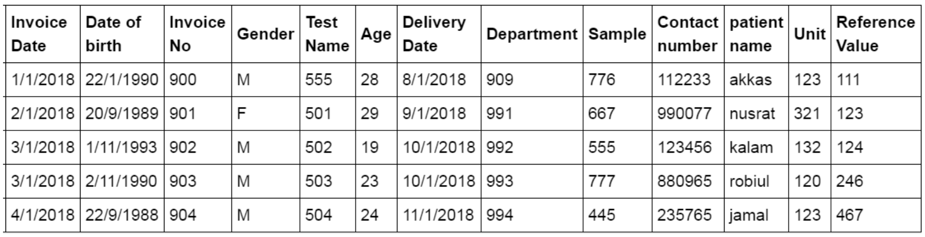
data conflicts, "K"(obviously) and the scoring capacity to rate the significance of each segment with the objective variable.

3. Recursive feature elimination (RFE) with random forest
    - It is a grasping improvement for discovering the top performing subset of features.
    - RFE depends on the plan to more than once manufacture a model, and pick either the finest or most noticeably bad performing feature, setting the features aside and after that reiterating the procedure with whatever is [12] left of the lineaments. This procedure is associated until all aspects in the dataset are depleted. Features are then situated by when they were discarded.
    - Essentially discard a particular number or a particular level of the minimum situating features in the model and retrains. This makes the supposition that takes features are not critical, regardless, so they can be discarded. It, by then continues taking out features in that limit, to the point that your stop premise is come to.
4. Recursive feature elimination with cross-validation and random forest classification
    - Cross-validation is a framework to survey perceptive models by dividing the first precedent into a planning set to set up the model, and a test set to evaluate it. In k-fold cross-validation, the principal model is subjectively separated into k measure up to estimate subsamples..

5 Results and Discussion

Firstly, we have collected the data from the Dataset and showing the first five rows from the datasets (see Table II). Than showing which types of data are available in. the Dataset (see Fig.3). Next step we are showing only the object columns (see Table III). Then we convert the dataset from Categorical Values to numerical Values for easily find out the main feature from the dataset (see fig.4). In the first algorithm, we find out the effective feature and showing the output (see Table IV) by domain knowledge. In the second algorithm, we find out the effective feature and showing (see Fig.5) by each attribute scores. In the third algorithm (see Fig.6) and fourth algorithm (see Fig.7) output are randomly generated. Therefore, we run this code 100 times to find out the effective feature and showing the descending order result in Fig.6 and Fig.7. In final step is to find out the common feature from all of the output of the algorithm and showing with numbering (seefig.8).

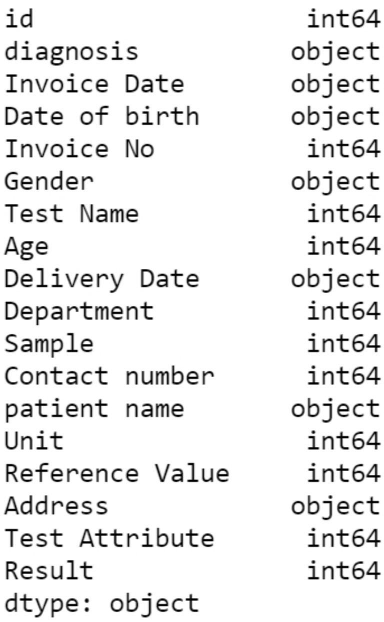
Table II. Show only first 5 rows from the dataset



Invoice Date	Date of birth	Invoice No	Gender	Test Name	Age	Delivery Date	Department	Sample	Contact number	patient name	Unit	Reference Value
1/1/2018	22/1/1990	900	M	555	28	8/1/2018	909	776	112233	akkas	123	111
2/1/2018	20/9/1989	901	F	501	29	9/1/2018	991	667	990077	nusrat	321	123
3/1/2018	1/11/1993	902	M	502	19	10/1/2018	992	555	123456	kalam	132	124
3/1/2018	2/11/1990	903	M	503	23	10/1/2018	993	777	880965	robiul	120	246
4/1/2018	22/9/1988	904	M	504	24	11/1/2018	994	445	235765	jamal	123	467

In this table (see Table II) shows only the first 5 rows from the dataset. In the dataset, there are 16 attributes.

After showing only the first 5 rows now we showing how many types of attributes (see Fig.3) in our dataset. Our dataset mixed with categorical and numerical values. Before using our dataset, all the data move into numerical.

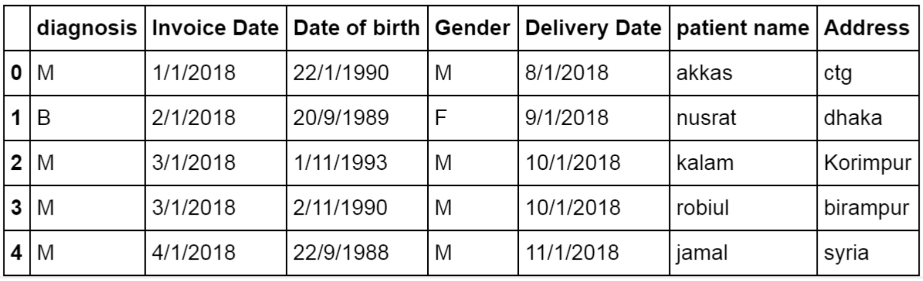


```
id                   int64
diagnosis           object
Invoice Date        object
Date of birth       object
Invoice No           int64
Gender              object
Test Name            int64
Age                  int64
Delivery Date       object
Department           int64
Sample               int64
Contact number       int64
patient name        object
Unit                 int64
Reference Value      int64
Address             object
Test Attribute       int64
Result               int64
dtype: object
```

Fig. 3. Data types in the dataset

Pandas incorporate a helpful select_dtypes perform that we will use to create a replacement data frame containing exclusively the item columns. In our dataset, there are several types of data. In Table III only showing the object columns attribute.

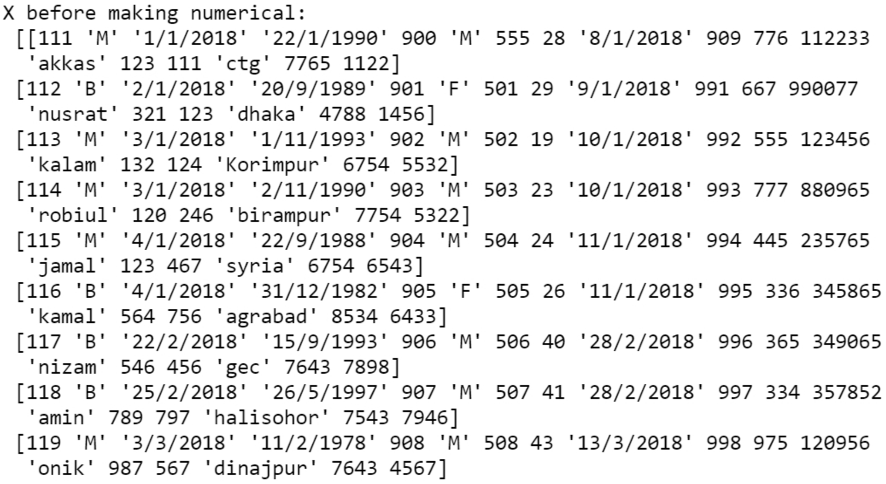
Table. III Shows only object columns



	diagnosis	Invoice Date	Date of birth	Gender	Delivery Date	patient name	Address
0	M	1/1/2018	22/1/1990	M	8/1/2018	akkas	ctg
1	B	2/1/2018	20/9/1989	F	9/1/2018	nusrat	dhaka
2	M	3/1/2018	1/11/1993	M	10/1/2018	kalam	Korimpur
3	M	3/1/2018	2/11/1990	M	10/1/2018	robiul	birampur
4	M	4/1/2018	22/9/1988	M	11/1/2018	jamal	syria

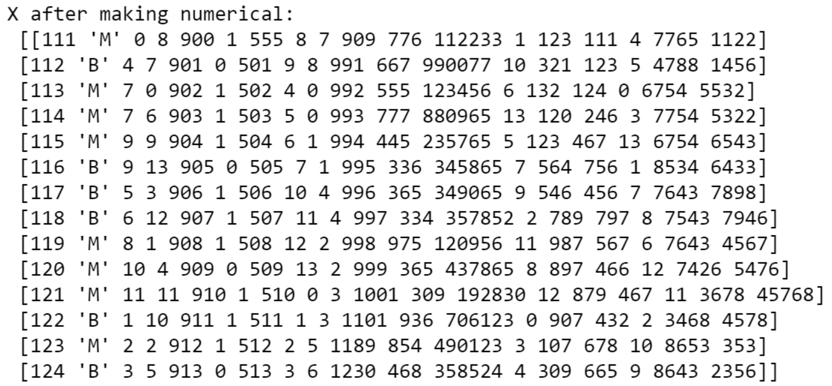
In our dataset, there are several types of that. But for implementation, we need to change all the data into arithmetical format. For this purpose, we Encoding all the Categorical Values and convert all the data into a numerical format (see Fig.4)

Before making numerical data:



```
X before making numerical:
 [[111 'M' '1/1/2018' '22/1/1990' 900 'M' 555 28 '8/1/2018' 909 776 112233
  'akkas' 123 111 'ctg' 7765 1122]
 [112 'B' '2/1/2018' '20/9/1989' 901 'F' 501 29 '9/1/2018' 991 667 990077
  'nusrat' 321 123 'dhaka' 4788 1456]
 [113 'M' '3/1/2018' '1/11/1993' 902 'M' 502 19 '10/1/2018' 992 555 123456
  'kalam' 132 124 'Korimpur' 6754 5532]
 [114 'M' '3/1/2018' '2/11/1990' 903 'M' 503 23 '10/1/2018' 993 777 880965
  'robiul' 120 246 'birampur' 7754 5322]
 [115 'M' '4/1/2018' '22/9/1988' 904 'M' 504 24 '11/1/2018' 994 445 235765
  'jamal' 123 467 'syria' 6754 6543]
 [116 'B' '4/1/2018' '31/12/1982' 905 'F' 505 26 '11/1/2018' 995 336 345865
  'kamal' 564 756 'agrabad' 8534 6433]
 [117 'B' '22/2/2018' '15/9/1993' 906 'M' 506 40 '28/2/2018' 996 365 349065
  'nizam' 546 456 'gec' 7643 7898]
 [118 'B' '25/2/2018' '26/5/1997' 907 'M' 507 41 '28/2/2018' 997 334 357852
  'amin' 789 797 'halisohor' 7543 7946]
 [119 'M' '3/3/2018' '11/2/1978' 908 'M' 508 43 '13/3/2018' 998 975 120956
  'onik' 987 567 'dinajpur' 7643 4567]
```

After making numerical data:



```
X after making numerical:
 [[111 'M' 0 8 900 1 555 8 7 909 776 112233 1 123 111 4 7765 1122]
 [112 'B' 4 7 901 0 501 9 8 991 667 990077 10 321 123 5 4788 1456]
 [113 'M' 7 0 902 1 502 4 0 992 555 123456 6 132 124 0 6754 5532]
 [114 'M' 7 6 903 1 503 5 0 993 777 880965 13 120 246 3 7754 5322]
 [115 'M' 9 9 904 1 504 6 1 994 445 235765 5 123 467 13 6754 6543]
 [116 'B' 9 13 905 0 505 7 1 995 336 345865 7 564 756 1 8534 6433]
 [117 'B' 5 3 906 1 506 10 4 996 365 349065 9 546 456 7 7643 7898]
 [118 'B' 6 12 907 1 507 11 4 997 334 357852 2 789 797 8 7543 7946]
 [119 'M' 8 1 908 1 508 12 2 998 975 120956 11 987 567 6 7643 4567]
 [120 'M' 10 4 909 0 509 13 2 999 365 437865 8 897 466 12 7426 5476]
 [121 'M' 11 11 910 1 510 0 3 1001 309 192830 12 879 467 11 3678 45768]
 [122 'B' 1 10 911 1 511 1 3 1101 936 706123 0 907 432 2 3468 4578]
 [123 'M' 2 2 912 1 512 2 5 1189 854 490123 3 107 678 10 8653 353]
 [124 'B' 3 5 913 0 513 3 6 1230 468 358524 4 309 665 9 8643 2356]]
```

Fig. 4. Convert all the data into a numerical format

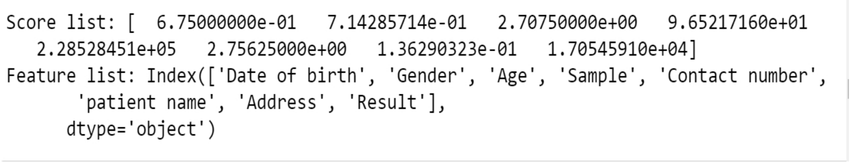
Now, our dataset only has numerical format data and ready to execute. Using this dataset now Implement this 4 random algorithm to find out the efficient features:

1. The first algorithm used for find out the effective feature by domain knowledge and showing the output result in Table IV.

Table IV. Shows the selected Feature

	Date of birth	Gender	Age	Sample	Contact number	patient name	Address	Result
0	8	1	8	776	112233	1	4	1122
1	7	0	9	667	990077	10	5	1456
2	0	1	4	555	123456	6	0	5532
3	6	1	5	777	880965	13	3	5322
4	9	1	6	445	235765	5	13	6543

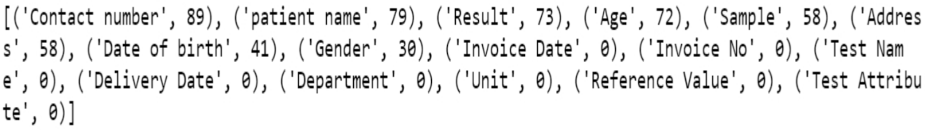
2. The second algorithm used for find out the effective feature by each attribute scores (seeFig.5).



```
Score list: [  6.75000000e-01   7.14285714e-01   2.70750000e+00   9.65217160e+01
   2.28528451e+05   2.75625000e+00   1.36290323e-01   1.70545910e+04]
Feature list: Index(['Date of birth', 'Gender', 'Age', 'Sample', 'Contact number',
       'patient name', 'Address', 'Result'],
      dtype='object')
```

Fig. 5. Univariate feature selection Output

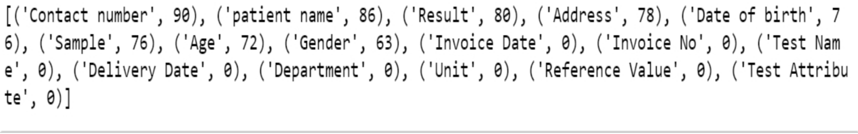
3. In the third algorithm, the output is randomly generated. So, run this code 100 times to find out the effective feature and showing the descending order result inFig.6.



```
[('Contact number', 89), ('patient name', 79), ('Result', 73), ('Age', 72), ('Sample', 58), ('Addres
s', 58), ('Date of birth', 41), ('Gender', 30), ('Invoice Date', 0), ('Invoice No', 0), ('Test Nam
e', 0), ('Delivery Date', 0), ('Department', 0), ('Unit', 0), ('Reference Value', 0), ('Test Attribu
te', 0)]
```

Fig.6. Recursive feature elimination (RFE)

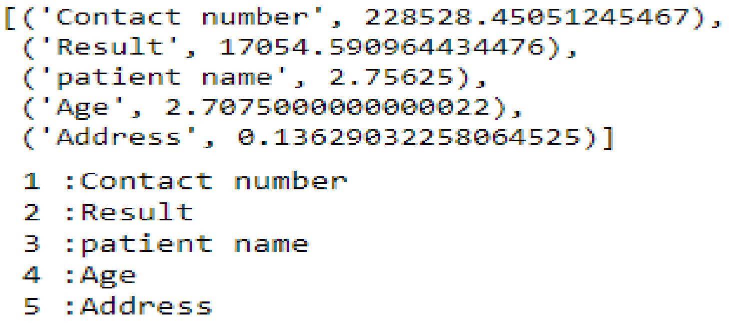
4. In the fourth algorithm, this algorithm also generates the result randomly. So, run this code 100 times to find out the effective feature and showing the descending order result in Fig.7.



```
[('Contact number', 90), ('patient name', 86), ('Result', 80), ('Address', 78), ('Date of birth', 7
6), ('Sample', 76), ('Age', 72), ('Gender', 63), ('Invoice Date', 0), ('Invoice No', 0), ('Test Nam
e', 0), ('Delivery Date', 0), ('Department', 0), ('Unit', 0), ('Reference Value', 0), ('Test Attribu
te', 0)]
```

Fig.7. Proper Feature selection process.

In final step is to find out the common feature from all of the output of the algorithm and showing with numbering (see fig.8).



```
[('Contact number', 228528.45051245467),
 ('Result', 17054.590964434476),
 ('patient name', 2.75625),
 ('Age', 2.7075000000000022),
 ('Address', 0.13629032258064525)]

 1 :Contact number
 2 :Result
 3 :patient name
 4 :Age
 5 :Address
```

Fig.8. numbering of Proper Feature selection process

6 Conclusions

Selection of a proper subset of features is very important in many types of research such as data integration and machine learning. It is always difficult to choose which features or attributes should we use for data integration for data warehousing. It requires deep knowledge of the problem domain. Automatic feature assortment is the procedure of choosing a subset of applicable features automatically for later use. In the future, we will add more security issue to increase efficiency. We will use more healthcare dataset. When the dataset will increase that time security and efficiency issue is more important. We will build and update our algorithm to ensure security and efficiency. We will also enhance the noise reduction technique.

References

[1] Shen, F., et al. (2015). Bearing fault diagnosis based on SVD feature extraction and transfer learning classification. 2015 Prognostics and System Health Management Conference (PHM), IEEE

[2] Khan, Shahidul & Latiful Haque, Abu. (2016). Towards Development of National Health Data Warehouse for Knowledge Discovery. 10.1007/978-3-319-23258-4_36

[3] Acharya, A. and D. Sinha (2014). "Application of feature selection methods in educational data mining." International Journal of Computer Applications 103(2)

[4] Bidgoli, A.-M. and M. N. Parsa (2012). "A hybrid feature selection by resampling, chi squared and consistency evaluation techniques." World Academy of Science, Engineering and Technology 68: 276-285

[5] Singh, B., et al. (2014). "A feature subset selection technique for high dimensional data using symmetric uncertainty." Journal of Data Analysis and Information Processing 2(04): 95. Ramaswami, M., Bhaskaran, R.: A Study on Feature Selection Techniques in Educational Data Mining. J.Comput. 1, 7– 11(2009).

[6] Jishan, S. T., et al. (2015). "Improving accuracy of students' final grade prediction model using optimal equal width binning and synthetic minority over-sampling technique." Decision Analytics 2(1): 1

[7] Tang, J., et al. (2014). "Feature selection for classification: A review." Data classification: Algorithms and applications: 37

[8] Kandel, S., et al. (2011). "Research directions in data wrangling: Visualizations and transformations for usable and credible data." Information Visualization **10**(4): 271-288.

[9] Lin, T. Y. and N. Cercone (2012). Rough sets and data mining: Analysis of imprecise data, Springer Science & Business Media

[10] Chuang, L.-Y., et al. (2016). "A hybrid both filter and wrapper feature selection method for microarray classification." arXiv preprint arXiv:1612.08669

[11] Ghaemidizaji, Manizheh & Feizi Derakhshi, Mohammad Reza. (2014). Classifying Different Feature Selection Algorithms Based on the Search Strategies

[12] Ong, T. C., et al. (2017). "Dynamic-ETL: a hybrid approach for health data extraction, transformation and loading." BMC medical informatics and decision making 17(1): 134

[13] Khan, S. I. and A. S. M. L. Hoque (2015). Development of national health data warehouse Bangladesh: Privacy issues and a practical solution. 2015 18th International Conference on Computer and Information Technology (ICCIT), IEEE

[14] M. Badiuzzaman Biplob, G. A. Sheraji and S. I. Khan, "Comparison of Different Extraction Transformation and Loading Tools for Data Warehousing," 2018 International Conference on Innovations in Science, Engineering and Technology (ICISET), Chittagong, Bangladesh, 2018, pp. 262-267