International Journal of Engineering and Advanced Technology (IJEAT)
ISSN: 2249-8958 (Online), Volume-9 Issue-3, February 2020

Generation of Association Rules of Data Mining for Lung Cancer by Air Pollution

S. Kanageswari, D. Gladis

Abstract: ***Revelation to adverse air pollutants attributed harmful effects in humans health. This research targets to evaluate the influence of atmospheric pollutants via determining the number of hospitalization underlying pulmonary complication in Chennai, Tamil Nadu. This tropical metropolitan city and also capital of Tamil Nadu have recently endured with the atmospheric pollutants. Due to rapid urbanization, followed by installation of numerous industries over the years have gradually affected the air quality. Chennai has respiratory illness in maximum record owing to atmospheric pollutants. The atmospheric pollutants and its impact on well-being could be due to pollutant's ability in inducing oxidative stress, allergy and irritation, and it is reasonable that high points for air pollutants is producing hospitalization in great number. In this paper, a efficacious and novel study utilizing data mining approach involving 'suggestion rules' had imparted, wherein its capability to search for an fundamental linking among qualities with greater database and the capacity to handle inexact database that frequently happens under real world scenario which appeared rapidly problematic. A detection of association dealings, regular designs or connections between items set or components in databases is association rules mining. Association rules are very beneficial in atmospheric pollutants and healthcare database because they deal prospect to lead smart analysis and produce valuable data also frame important data bases rapidly and routinely, so that progress effective plans to minimize health contact to the atmospheric pollutants. Data completed pre-processing phase to assist condition of demonstrating procedure. With respect to conclusion, association rules mining had performed by Apriori, Eclat and FP growth algorithm the results showed that the latter was much accurate and consumes lesser time.***

Keywords: ***Pulmonary complication, association rule, atmospheric pollutants, apriori, data mining, Eclat, FP growth***

I. INTRODUCTION

Various health problems begins from slight eye irritations followed by upper respiratory symptoms, CVD, chronic respiratory infections and lung cancer caused because of atmospheric pollutants. These might consequence in hospital admittance and sometimes death [1]. Through environment and health, particulate matter (PM) concentration have serious impacts. Therefore, assessment of PM trend become important globally, explicitly in urban regions. PM monitoring made as a compulsory one by several government forms. It built separate standards to lessen the effects. Compared with National Ambient Air Quality Standards of India, Annual PM Concentration is higher as per the reports of major cities. Health impacts caused by deposited PM, at lower concentrations. This outcomes in short-range and long-standing effects like hospital admissions (Luong et al. 2017) , chronic obstructive pulmonary disease (COPD), critical lower respiratory infection (ALRI), pneumonia, stroke reduced heart rate variability, CVD, DNA damage, low birth weight, increased likelihood of suicide, and ultimately mortality. Business and advertising administrations may be ahead of healthcare in applying data mining for rendering knowledge from data. This is quickly changing. Successful of application mining had implemented in healthcare arena. [2]

In 2015, Ruhul Amin Dicken [3] suggested a scheme. That scheme connects patients and pollutants of admittance in the hospital and examines the cause ahead for exponential rise in disease reported in Bangladesh hospitals. For clustering altered atmospheric pollutants in various periods of Bangladesh, this scheme found k-means clustering approach and CART way to organize patients conferring to various level of admittance. Simpson et al. [4] proposed a connection among air pollutants and its association over respiratory illness among juvenile population. The study was conducted with utilizing hospital admission dataset for the period 1998-2001 they focused on three years children, the datasets were from five most important cities of Australia and two cities of New Zealand. In US, the Environmental Protection Agency (EPA) is accountable for generating, changing, and implementing ethics for air quality in order to defend the public from confrontational health effects.

In lung cancer prediction, a system for Conclusion Support Vector Machine and Naïve Bayes through techniques of current attribute choice is employed by Bin Liu in 2016. A technique similar Random Forest and Naive Bayes gives improved outcome in lung cancer prediction, for which Decision tree was employed. For aggregating the exactness of the perfect collective technique is to be preferred. Rather allocating weights by hand is not a best approach. Consequently, in current ages, iDHS-EL and iDNA-KACC-EI approaches through collective tactic are used to a discover weight on behalf of fusion procedure [5]. IRSpot-EL technique of collective used for fusion analyses the distance by applying sympathy propagation algorithm [6]. Clustering is a method of extrication dataset into subgroups conferring to single feature. As regards Lung Cancer, it separates the dataset into required and non-required dataset. To discover common designs of dataset, Apriori [7] and Decision Tree algorithm and Elad, [8] are essentially used. These algorithms are effective and accurate in finding the FP.

Revised Manuscript Received on February 05, 2020.

* Correspondence Author

S. Kanageswari*, Research Scholar, Computer Science and Engineering Bharathiar University, Coimbatore, Tamilnadu, India. skanageswari@gmail.com

Dr. D. Gladis, Principal, Computer Science and Engineering, Bharathi womens college (autonomous), Tamilnadu, India

Frequent patterns, refers to the data that frequently repeated in the dataset. There are some frequent patterns which are accountable to lung cancer; through this pattern implementation of prediction system for Lung Cancer is carried out.

V. Krishnaia, [9] Knowledge Discovery in Databases (KDD), which contains data mining techniques, has develop a popular research tool for medical researchers to recognize and exploit arrangements and connection between huge number of variables, and through them which is capable to estimate consequence of a disease using in historical cases and its storage is achieved within datasets. This research outline numerous analysis and methodological articles pertaining to diagnosis involving lung cancer. It provides a summary of present research being accepted on numerous lung cancer datasets using the data mining techniques to increase the specificity.

In 2018 initial nation-wide lung cancer risk calculation related by out-of-doors PM2.5 experience in France considering dissimilar dose-response functions [10]. The best key manageable sources of lung cancer among non-smokers is PM2.5. Various policy act emphasized to decrease PM2.5 attentions in France having huge capacity in minimizing lung cancer cases. Moreover, developments in the quality and attention concerning with atmospheric pollutants data followed by atmospheric pollutants mixture are regarded critical for examination when monitoring impacts on future health.

In this research we inspected whether monitoring the temporal environmental exposure profile of a Chennai region is correlated with the rise in cancer incident. As indicative we use risk underlying lung cancer and its association to prolonged disclosure to various PM. This research uses the association rule to observe the effects of air pollution contact on patients of lung cancer. In case-crossover design which facilitates in considering the properties of exact exposures that can also study multiple exposures then relations among exposures. This method had used to the research of severe effects of environmental experiences, particularly air pollution. The study is carried out with Apriori algorithm, ECLAT algorithm and Frequent-Pattern growth algorithm, comparative research is prepared among classical frequent pattern mining algorithms which use the patient set generation and test (Apriori algorithm) and the algorithm without candidate set generation (FP growth algorithm). To find out the frequent item set, Apriori algorithm used and the subset should also be frequent. To creates candidate item set and similarly test whether it is frequent or not Apriori algorithm used. From huge database Frequent-Pattern growth technique uses pattern fragment development which is essential to mine FP. Based on depth-first algorithm, Eclat algorithm is a vertical data representation.

II. PROPOSED METHODOLOGY

To maintain the clinician in creating improved evaluation, data mining tools established for effective analysis. Present study involve data collection via Hospital Information System (HIS) comprising adequate particulars on patients namely: patient's name, disease, location, age, district, and date from laboratories and it retains on rising year after year. Using composed data from HIS, this study can discover the common disease with the support of association techniques. This study work benefits to mine nearby DATA the common diseases through benefit of python tool applied completed training data set.

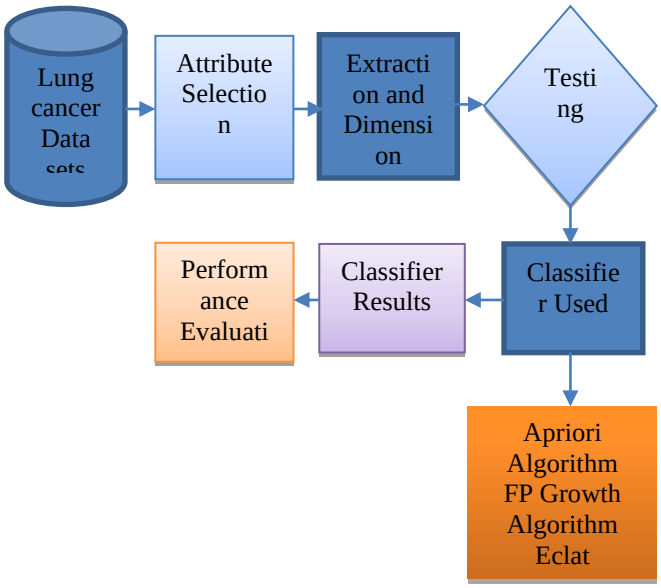

Figure 1: Block diagram of overall process

APRIORI ALGORITHM

It is considered as greatest classical and significant algorithm for mining frequent item sets [11]. To discover all several itemsets under given database (DB), Apriori algorithm used. Using its principle with any subset of a repeated itemset needed to be frequent. For Instance: if {XY} defines as frequent itemset, as both {A} and {B} must be frequent itemsets. The significance pertaining to Apriori algorithm involve generation on numerous documentations of database. Through search space, it uses an iterative approach also known as Breadth-first search (level-wise search) where k-itemsets are used to explore (k+1)-itemsets. Upon commencement, set involving frequent 1-itemsets is discovered. The set comprises one item, which accomplish the provision threshold, is denoted by L1. In individually subsequent pass, we instigate with a seed set of itemsets originate to be great in earlier pass. For producing fresh possibly great itemsets, seed set is used. It is known also known as candidate itemsets. Through data it count authentic maintenance for these candidate itemsets passes. We control which of the candidate itemsets are significantly great (frequent) at the end of the pass. Also established the seed for subsequent pass. Consequently, L1 employed in discovering L2, the set of frequent 2-itemsets, and it helped to discover L3, and therefore on, till there is no further frequent k-itemsets may be originate [12].

FP Growth algorithm

This algorithm [13, 14], however being a standard ARM algorithm, and is effective for mining large sets of data with FPs (i.e., a large amount of "large" items). Its competence falsehoods in the compressed and whole method it symbolizes the complete set of connections and designs in a tree-like form, which removes the requirement to make candidate items sets.

Though, for enormously enormous datasets, the frequent-tree structure can also surpass the key memory, producing YO overheads. We suggest the Partitioned Frequent-Pattern Growth algorithm, which is created on partitioning the transaction database, and processing each of the partitions in parallel. This method enables the efficient processing of huge datasets with long patterns.

Eclat algorithm

Based on depth-first algorithm, it is a vertical data representation. To increase the speed of frequent item set generation, it uses inverted table. On the other hand, resulting in a huge amount of candidate set, deletion of candidate set is not executed also influences the algorithm efficiency and it is one of the major disadvantage.

Dataset used

The dataset was obtained for 2015 from the Tamil Nadu air Pollution Control Board (TNPCB) with the attributes of PM10, SO2 and NO2. The total hospital admissions are taken from the survey reports of lung cancer owing to the air pollution. The four attributes are used for an input to predict attribute. And the last attribute was the patients target attributes. All attributes contain distinct values. The association rule accepts the datasets in the discrete form. Table 1 signifies the various air quality parameters. Table 2 signifies the various frequency of the data set with respect to the major areas of Chennai.

Table 1: Attributes for patients and quality of air parameters

No	Attribute	Measurement Unit	Data Notation
1	Site	not relevant	Station
3	PM_{10}	Microgram/ Cubic Metre	Particulate Matter
4	SO_2	Microgram/ Cubic Metre	Sulphur Dioxide
5	NO_2	Microgram/ Cubic Metre	Nitrogen Dioxide

Table 2: Raw Datasets

Site	PM_{10}	So_2	No_2	Patients
AnnaNagar	113	14	32	7
TNagar	97	16	51	11
Adyar	75	05	16	14
Kilpauk	102	10	51	12
AnnaNagar	113	14	32	9
TNagar	97	16	51	16
Adyar	75	05	16	21
Kilpauk	102	10	51	25
AnnaNagar	113	14	32	16

Data Discretization

Particularly for huge amount of datasets, attributes that are not complete, data that misses few missing value, discretization is sufficient. To discretize the data Han and Kamber [15], the equal frequency binning technique was used. Data discretization on every attributes are presented in table 3.

Table 3: Data discretization

PM_{10}	So_2	No_2	Patients
Normal	High	High	Moderate
Normal	Low	High	High
High	Low	Normal	High
normal	Low	Normal	Moderate
High	Low	High	High
High	Normal	Normal	Moderate
High	Normal	High	High
Normal	Normal	Normal	Moderate
Low	Normal	Low	Normal
Low	High	Low	Normal
Low	High	Low	Normal
Low	High	Low	Normal
Low	High	High	Moderate
Low	Low	Normal	Normal

Algorithm used

Implementation

The following algorithms are used.

1. Apriori
2. FP Growth
3. ECLAT

Apriori Algorithm

Based on the information, the term of Apriori algorithm is generated. It uses prior data of frequent itemset properties which is that all nonempty subsets of a frequent itemset should also be frequent [16]. Finding the frequent itemsets is a main strategy.

Algorithm methods:

- Fix the least support and confidence conferring to user definition.
- If support values are lesser than the minimum support, form the candidate 1-itemsets then produce the frequent 1-itemsets by cropping some candidate 1-itemsets
- To form the candidate 2-itemsets, connect the frequent 1-itemsets with each other. To generate the frequent 2-itemsets, crop some infrequent itemsets from the candidate 2-itemsets.
- Until no more candidate itemsets generated, follow the previous method.

Python pseudocode- Apriori Algorithm

```
import numpy as np
import pandas as pd
from efficient_apriori import apriori
df=pd.read_excel('air_input.xlsx')
final_df=df[['RSPM','SO2','Nox','Patients_Status']]
records=[]
for i in range(len(final_df)):
records.append([str(final_df.values[i,j]) for j in range(len(final_df.columns.tolist()))])
itemsets, rules=apriori(records, min_support=0.1,min_confidence=0.1)
rules_rhs = filter(lambda rule: len(rule.lhs) == 1 and len(rule.rhs) == 1, rules)
for rule in sorted(rules_rhs, key=lambda rule: rule.lift):
```

```
if
(rule.lhs==('Patients_high',))|(rule.rhs==('Patients_high',)):
 print(rule)
```

FP Growth

By Han,Pei and Yin the Frequent pattern growth was invented in 2000 to waive candidate generation completely [17]. It is accomplished by using a trie to store the actual baskets, instead of storing candidates such as Apriori and Eclat do. The Aprori algorithm is more horizontal, breadth-first.

Python pseudocode- FP Growth Algorithm

```
import time
from datetime import timedelta
start_time = time.monotonic()
import numpy as np
import pandas as pd
import pyfpgrowth
df=pd.read_excel('air_input.xlsx')
final_df=df[['RSPM','SO2','Nox','Patients_Status']]
records=[]
for i in range(len(final_df)):
records.append([str(final_df.values[i,j]) for j in
range(len(final_df.columns.tolist()))])
patterns = pyfpgrowth.find_frequent_patterns(records, 2)
rules = pyfpgrowth.generate_association_rules(patterns,
0.1)
filtered_dict = {k:v for (k,v) in rules.items() if
v[0]==('Patients_high',)}
end_time = time.monotonic()
print(timedelta(seconds=end_time - start_time))
```

ECLAT

In 1997 by Zaki, Parthasarathy, Ogihara, and Li Eclat was introduced [18]. Equivalence Class Clustering and bottom up Lattice Traversal (ECLAT). The key variance among Eclat and Apriori is that Eclat abandons Apriori's breadth-first analysis for a recursive search of depth-first. In 2003 by Goethals the Eclat is described. [19]. Based on depth-first algorithm the Eclat algorithm is a vertical data demonstration. The Eclat algorithm uses an inverted table to continuously increase the speed of item set generation. Though, the complication is that the removal of the patient set is not executed, which resulting in a huge amount of patient sets, and it affects the algorithm efficiency.

Python pseudocode- Eclat Algorithm

```
import time
from datetime import timedelta
start_time = time.monotonic()
import numpy as np
import pandas as pd
df=pd.read_excel('air_input.xlsx')
final_df=df[['RSPM','SO2','Nox','Patients_Status']]
final_df=final_df[final_df['Patients_Status'] ==
'Patients_high']
final_df = final_df[['RSPM','SO2','Nox']]
records=[]
for i in range(len(final_df)):
records.append([str(final_df.values[i,j]) for j in
range(len(final_df.columns.tolist()))])
from fim import eclat
rules = eclat(tracts = records, zmin = 1)
rules.sort(key = lambda x: x[1], reverse = True)
end_time = time.monotonic()
print(timedelta(seconds=end_time - start_time))
```

III. EVALUATION PARAMETERS

A comparison framework had developed to allow the flexible comparison of the various algorithms that conform to the defined performance of algorithm. The performance comparison is carried respecting to run time or execution time and respecting to the overall accuracy. Table 4 presents the run time values of the various algorithm and table 5 presents the overall correctness of the various algorithm.

Run time

The below table shows the run time of the algorithms used in the study. On comparing the timing the highest time is consumed by the Apriori algorithm with 31milliseconds carried by Eclat algorithm by 18 milliseconds, the least time is considered with the FP growth algorithm. The main difference in the timing is due to the number of scans to generate the patient's sets. Apriori algorithm involves multiple scan for generating the data set, similarly Eclat algorithms scan continuously to generate the patient data, however only two times the database of FP growth algorithm is scanned. This results in less performance time on using FP growth algorithm on comparison with other two algorithms.

Table 4: Run time

Algorithm	Run Time
Apriori Algorithm	0:00:00.031000
FP Growth Algorithm	0:00:00.015000
Eclat Algorithm	0:00:00.018000

Accuracy

The table 5 and the figure 2 displays the overall accuracy of the various algorithms used in the study. The highest performance with respect to the accuracy is noticed with respect to the FP growth algorithm and the least accuracy is noticed with respect to the Apriori algorithm.

Table 5: Overall performance of Algorithm employed (in terms of accuracy)

Algorithm	Accuracy
Apriori Algorithm	78.24%
FP Growth Algorithm	83.21%
Eclat Algorithm	81.12%

IV. RESULTS ANALYSIS

The paper is contributed for the deceased and surviving Lung cancer patients. Also there is a greater likelihood for people who are susceptible to fall under this deadly disease that is impacted primly due to atmospheric pollutants. This is the first attempt using association rules in trying to understand the formation of air pollution from the effects of respiratory disease,

Thus solving the environmental issue. The acquired knowledge model can be used as a decision support system to acquire sets of knowledge that are useful in the context of preventing the increased risk of harmful air pollutants that cause lung cancer. The ambient quality of air is assessed from central pollution control board (CPCD), TNPCD ambient of data air quality. Air pollutants that are gasped have severe influence on human health affection the lung and respiratory system. It is also occupied up using the blood and pumped all-round the body. Data are pre-processed and data can be additionally processed by data mining tools and policy makers can be given appropriate decision support. Meanwhile the government has permitted several measures to deal with this problem. The prediction of air pollution in an urban area using data mining research studies will be carried forward with data obtained from a monitoring station in Tamil Nadu, India, which serves as a significant reference for policy markers in expressing upcoming policies.

The accuracy and the overall run time performance are compared and provided in the table 6. The graphical representation of the same are provided in the figure.

Table 6: Overall comparison

Algorithm	Accuracy	Run time
Apriori Algorithm	78.24%	0:00:00.031000
FP Growth Algorithm	83.21%	0:00:00.015000
Eclat Algorithm	81.12%	0:00:00.018000

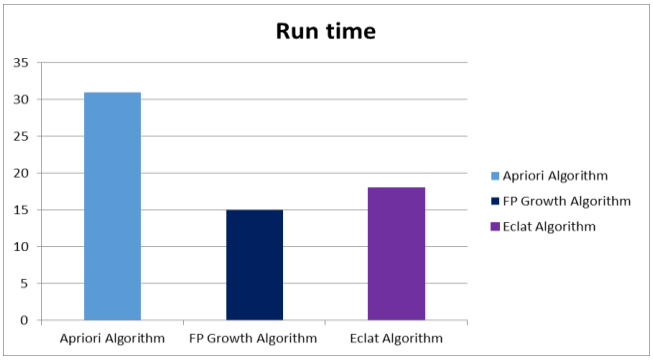

Figure 2: Comparisons of algorithms based on run time

The above figure shows the graphical presentation of the run time of the algorithms in millisecond. The maximum value of 31 millisecond is seen relating to the Apriori algorithm and the least time of 15millisecond is seen relating to the FP growth algorithm.

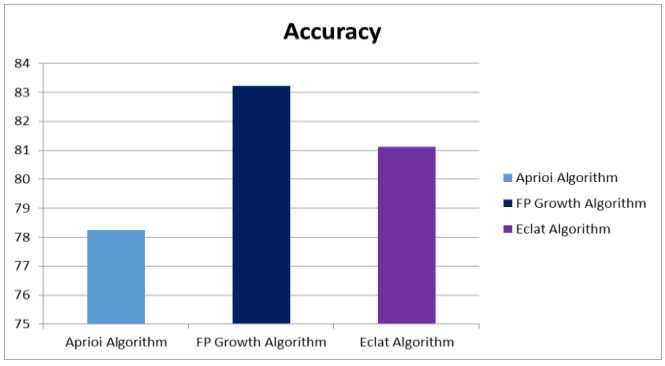

Figure 3: Comparisons of algorithms based on Accuracy

The figure shows the graphical data indicating the overall accuracy of the algorithms used in the study. Highest accuracy is achieved as 83.21% with the use of FP growth algorithm and the next with the value of 81.12% with the use of Eclat algorithm and the least value of 78.24% with the use of Apriori algorithm. Table 7, provides a consolidated findings, followed by future studies to be employed in highly polluted districts, other than Chennai.

Table 7: Important factors/ pollutants, major sites and future works to emphasise on Data mining approaches for lung cancer determination in Tamil Nadu

Factors/ pollutants	Major sites in Chennai city observed	Future studies to emphasise on data mining based surveying on Lung cancer patients
PM_{10}	Kilpauk	Thoothukudi
SO_2	Adyar	Trichy
NO_2	Annanagar, T-Nagar	Coimbatore

The association rule mining technique called Apriori algorithm, frequent growth pattern (FP growth) and Eclat algorithm. The difference between apriori, FP growth and the Eclat is the runtime. The main reason for this the number scans performed in each algorithm not same. FP growth is effectual than apriori and Eclat algorithms with respect to the both accuracy and the run time. This study has given the valuable contribution to the air pollution management and used to understand the different air pollutant, which influence to the respiratory illness.

V. CONCLUSION

Thus from the results it's clear that FP growth algorithm perform well when compared to other algorithms. From this study it can be concluded that FP growth algorithm will be the best classifier in testing of lung cancer data sets with high accuracy.

DISCUSSION

In decision, associations through risk for lung cancer were found with numerous PM elements from different foundations; the strongest relatives were seen for participants. When participants did not variation their address through follow-up. Since strengths and confines, this research specifies that the suggestion among PM in air pollution and lung cancer can be qualified to several PM components and sources.

REFERENCES

1. Zheng, M. 2011. Hong Kong: Particulate air pollution and health impacts. *Encyclopedia Environmental Health*, pp. 56-61.
2. Miller, R. A. (1994). Medical diagnostic decision support systems—past, present, and future: a threaded bibliography and brief commentary. *Journal of the American Medical Informatics Association*, *1*(1), 8-27.
3. Dicken, R. A., Rubby, S. M. F., Naz, S., Khaled, A. A., Rahman, S. A., Rahman, S., & Rahman, R. M. (2015, June). Analysis and classification of respiratory health risks with respect to air pollution levels. In *2015 IEEE/ACIS 16th International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD)* (pp. 1-6). IEEE.
4. Barnett, A. G., Williams, G. M., Schwartz, J., Neller, A.

H., Best, T. L., Petroeschevsky, A. L., & Simpson, R. W. (2005). Air pollution and child respiratory health: a case-crossover study in Australia and New Zealand. *American journal of respiratory and critical care medicine*, *171*(11), 1272-1278.

5. Liu, B., Wang, S., Long, R., & Chou, K. C. (2016). iRSpot-EL: identify recombination spots with an ensemble learning approach. *Bioinformatics*, *33*(1), 35-41.
6. Pantola, P., Bala, A., & Rana, P. S. (2015, August). Consensus based ensemble model for spam detection. In *2015 international conference on advances in computing, communications and informatics (ICACCI)* (pp. 1724-1727). IEEE.
7. Lan, C., Liu, Y., & Tang, Z. (2010). Improvement of aprioritid algorithm for mining frequent items [J]. *Computer Applications And Software*, *27*, 234-6.
8. Ben-Haim, Y., & Tom-Tov, E. (2010). A streaming parallel decision tree algorithm. *Journal of Machine Learning Research*, *11*(Feb), 849-872.
9. Krishnaiah, V., Narsimha, D. G., & Chandra, D. N. S. (2013). Diagnosis of lung cancer prediction system using data mining classification techniques. *International Journal of Computer Science and Information Technologies*, *4*(1), 39-45.
10. Kulhanova, I., Morelli, X., Le Tertre, A., Loomis, D., Charbotel, B., Medina, S., ... & Soerjomataram, I. (2018). The fraction of lung cancer incidence attributable to fine particulate air pollution in France: Impact of spatial resolution of air pollution models. *Environment international*, *121*, 1079-1086.
11. Agrawal, R., Imielinski, T., & Swami, A. (1993) Mining association between sets of items in massive database. In: International proceedings of the ACM-SIGMOD international conference on management of data (pp. 207–216).
12. Rao, S., & Gupta, P. (2012). Implementing Improved Algorithm Over APRIORI Data Mining Association Rule Algorithm 1.
13. Hipp, J., Güntzer, U., & Nakhaeizadeh, G. (2000). Algorithms for association rule mining- a general survey and comparison. *SIGKDD explorations*, *2*(1), 58-64.
14. Han, J., Pei, J., & Yin, Y. (2000, May). Mining frequent patterns without candidate generation. In *ACM sigmod record* (Vol. 29, No. 2, pp. 1-12). ACM.
15. Han, J., Kamber, M., & Tung, A. K. (2001). Spatial clustering methods in data mining. *Geographic data mining and knowledge discovery*, 188-217.
16. Hart, J., & Kamber, M. (2001). Data mining: concepts and techniques. *M or an Kaufmann Publishers*, *200*(1), 223-259.
17. Han, J., Pei, J., & Yin, Y. (2000, May). Mining frequent patterns without candidate generation. In *ACM sigmod record* (Vol. 29, No. 2, pp. 1-12). ACM.
18. Zaki, M. J., Parthasarathy, S., Ogihara, M., & Li, W. (1997). Parallel algorithms for discovery of association rules. *Data mining and knowledge discovery*, *1*(4), 343-373.
19. Goethals, B. (2003). Survey on frequent pattern mining. *Univ. of Helsinki*, *19*, 840-852.