International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 11 Issue 05, May-2022

# Pattern Mining of Hospitalization Data of Covid-19 Patients with Underlying Conditions

Nwagwu U[1], Ayinde A.Q[2], Isolagbenla K.O[3], Yusuf A.S[4]
[1]. Wichita State University Kansas City, USA
[2]. Hydropoint Data System, Petaluma, CA, USA
[3]. Southern New Hampshire University, Manchester, NH, USA
[4].New York Institute of Technology, Old Westbury, NYC, USA

***Abstract:*** **The covid-19 hospitalization rate is higher among 65years and above, since most of this individual have an underlying condition and with highest percentage of them living in assisted facilities. This research conducted a cluster relationship pattern mining between age, sex, underlying condition, and hospitalization status in five states in United States of America. Relationship between these data were evaluated before data were preprocessed. Over 1million data were preprocessed and summarized in Waikato Environment for Knowledge Analysis. Pattern recognition algorithms were applied to build a hospitalization cluster for a summarized data for the age group within this 1million population. The hospitalization patterns within this age bracket were analyzed.**

***Keyword*: *Pattern Mining, Covid 19, Knowledge Discovery, Algorithms, Hospitalization, Underlying Condition***

## 1.INTRODUCTION

The process of containing the spread and lowering the Covid-19 hospitalization rate has led the government to institute a variety of control measures via both government and the NGOs across the world. Pre data analyses were conducted based on an available public data and were correlated with data from the Centre for Disease Control (CDC). Evaluating the relationship between age group (65years and above) and underlying condition was tagged as factor 1 while relationship between age group (65years and above) and hospitalization status was tagged as factor 2. The data were summarized into Hospitalized Date, Number Hospitalized Per State and the Number Hospitalized by State Rolling Total. Five states namely Indiana, North Carolina, New York, Ohio and Pennsylvania hospitalization data were extracted from the master data, preprocessed using the constraints based sequential pattern mining to identify the frequent patterns in the hospitalization data

## 2.LITERATURE REVIEW

Constraint-based sequential pattern mining that rely on a multi-valued decision diagram (MDD) accommodate multiple items. Maintaining the integrity of the applicability off an MDD-based prefix-projection algorithm and compare its performance against a typical generate-and-check variant, as well as a state-of-the-art constraint-based sequential pattern mining algorithm [1] Sequential Pattern Mining (SPM) is a fundamental data mining task with a large array of applications in marketing, health care, finance, and bioinformatics, to name a few. Frequent patterns are used, e.g., to extract knowledge from data within decision support tools, to develop novel association rules, and to design more effective recommender systems [2]. Graphical representations of a database have been shown to be effective in item-set mining and SPM [3].

## 3.METHODOLOGY

This research adopted cross industry standard process for data mining (CRISP -DM). Data were preprocessed and summarized into clusters before partitioning into training and testing sets. For even calibration and data adjustment, 65percent of the data were used in the training and 35percent were used in testing using the explorer application of Waikato environment for knowledge analysis.

The determining variables based on this research data were *month*, state, county, race and ethnicity while determinant variables were age group, sex and hospitalization status. Assisted living/care giving homes population per county were calculated and was classified as high, normal and low.

Factor 1= 1 – a (b ∩ c) ……... (1)
Factor 2 = 1- a (c ∩ d) …… (2)

a= population size b=count of patients that are (65years and above) c = underlying condition d= number hospitalized per cluster.

Total of 1,200,000 dataset was extracted from the CDC website and Cross Industry Standard Process for Data Mining was adopted. The data were summarized to captured data extracted from the data source (CDC website). Data was summarized into 36,567 rows and 12 attributes. The cluster model on the Explorer platform were trained using the percentage split of 70percent for classes to cluster evaluation and 30percent for testing at different iteration. The output of the model after training and testing is in the snippet below with their cluster's instances.

Figure 1: Clustered Instances Analysis

| | |
|---|---|
| Cluster 0 | 2488 (23%) PA hospitalization is highest |
| Cluster 1 | 1643 (15%) NY hospitalization is highest |
| Cluster 2 | 3976 (36%) IN hospitalization is highest |
| Cluster 3 | 1526 (14%) OH hospitalization is highest |
| Cluster 4 | 1338 (12%) NC hospitalization is highest |

Figure 2: Model Cluster Analysis

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 11 Issue 05, May-2022

```
Number of clusters selected by cross validation: 5
Number of iterations performed: 7


                    Cluster
Attribute                0          1          2          3          4
                    (0.23)     (0.15)     (0.36)     (0.14)     (0.11)
=============================================================================
State
  NY                 5.8928 3861.4579     7.9152     7.9309     3.8032
  PA              5647.4387     1.2206      8.002     1.3829     7.9558
  NC                 7.9974      1.221 2367.2427     1.3826 2914.1563
  OH               212.2128     7.9978 2325.7279 3589.9728     8.0887
  IN                24.7024    80.0201 4512.6922     8.1961     6.3892
  [total]         5898.244 3951.9174 9221.5801 3608.8653 2940.3933
Hospitalization
  Yes              712.2948    249.027  370.7253  986.0602  670.8928
  No              1192.395  462.6967 8839.8204    8.0297    8.0582
  Missing            1.0435 3236.0103    3.3831 2605.5574    1.0057
  Unknown         3991.5107    3.1834    6.6514     8.218 2259.4366
  [total]          5897.244 3950.9174 9220.5801 3607.8653 2939.3933


Time taken to build model (percentage split) : 17.09 seconds
```

Table1: Model Metrics

| | Cluster0 | Cluster1 | Cluster2 | Cluster3 | Cluster4 |
|---|---|---|---|---|---|
| TPR | 0.78 | 0.72 | 0.56 | 0.67 | 0.98 |
| FPR | 0.43 | 0.57 | 0.32 | 0.77 | 0.18 |
| CL | 0.78 | 0.65 | 0.55 | 0.69 | 0.89 |

TPR = True Positive Rate FPR = False Positive Rate
CL = Convergence Level

## 4.RESULT DISCUSSION

The model formed a super cluster at cluster4 with the highest precision of 0.11. Hospitalization rate was on the average as evaluated by the by SPM. Factor 2 relationship predominated the pattern mining which make the model to converge at iteration 7 based on the sequential relationship between Factor 1 and Factor 2. From Table 1, the true positive rate and the convergence level values validated the relationship pattern between the underlying conditions and number hospitalized per cluster. The pattern analysis revealed that hospitalization rate at Cluster 0,2,3 and 4 for New York is low which is equivalent to the behavior exhibited by Pennsylvania hospitalization pattern from Cluster 1,2,3 and 4.

The analysis also revealed that Ohio, North Carolina and Indiana patten of hospitalization are similar. This similarity is due to the lower relationship between underlying conditions and number hospitalized per cluster. The SPM revealed that Factor 1 is greater than Factor 2 that is, Factor 1 dominated the cluster distribution by 65percent and even represented 80percent of the summarized data used in this research.

The cross validation and percentage split form the calibrating method that were adopted before the model can learn from the historical data. While iterations were performed at intervals as shown by Fig 1.

## REFERENCES

[1] Hosseininasab, A., Hoeve, W.-J. van, & Cire, A. A. (2019). Constraint-Based Sequential Pattern Mining with Decision Diagrams. Proceedings of the AAAI Conference on Artificial Intelligence, 33(01), 1495 – 1502 https://doi.org/10.1609/aaai.V33i01.33011495.

[2] Fournier-Viger, P.; Lin, J. C.-W.; Kiran, R. U.; Koh,Y. S.; and Thomas, R. 2017. A survey of sequential pattern mining. Data Science and Pattern Recognition 1(1):54–77.

[3] Han, J.; Pei, J.; Mortazavi-Asl, B.; Pinto, H.; Chen, Q.; Dayal, U.; and Hsu, M. 2001. Prefixspan: Mining sequential patterns efficiently by prefix-projected pattern growth. In proceedings of the 17th international conference on data engineering, 215–224.