Predicting Hospitals Hygiene Rate during COVID-19 Pandemic

Abdulrahman M. Qahtani[1], Bader M. Alouffi[2], Hosam Alhakami[3], Samah Abuayeid[4], Abdullah Baz[5]
Department of Computer Sciences, College of Computers and Information Technology, Taif University
Taif, Saudi Arabia[1,2]
Department of Computer Science, College of Computer and Information Systems, Umm Al-Qura University
Makkah, Saudi Arabia[3,4]
Department of Computer Engineering, College of Computer and Information Systems, Umm Al-Qura University
Makkah, Saudi Arabia[5]

***Abstract*—COVID-19 pandemic has reached global attention with the increasing cases in the whole world. Increasing awareness for the hygiene procedures between the hospital's staff, and the society became the main concern of the World Health Organization (WHO). However, the situation of COVID-19 Pandemic has encouraged many researchers in different fields to investigate to support the efforts offered by the hospitals and their health practitioners. The main aim of this research is to predict the hospital's hygiene rate during COVID-19 using COVID-19 Nursing Home Dataset. We have proposed a feature extraction, and comparing the results estimating from K-means clustering algorithm, and three classification algorithms: random forest, decision tree, and Naive Bayes, for predicting the hospital's hygiene rate during COVID-19. However, the results show that classification algorithms have addressed better performance than K-means clustering, in which Naive Bayes considered the best algorithm for achieving the research goal with accuracy value equal to 98.1%. AS a result the research has discovered that the hospitals that offered weekly amounts of personal protective equipment (PPE) have passed the personal quality test, which lead to a decrease in the number of COVID-19 cases between the hospital's staff.**

***Keywords*—COVID-19; machine learning; hospitals hygiene; World Health Organization (WHO); personal protective equipment; K-means clustering; Naive Bayes; random forest**

I. Introduction

Recently WHO has faced many challenges in increasing the global healthcare and Hygiene awareness to overcome COVID-19 pandemic. According to WHO COVID-19 is an infectious disease caused by a coronavirus, which started in December 2019 in the city of Wuhan in China [1]. On the other hand, hospitals and health practitioners are the most susceptible to infectious diseases. As a result, WHO has provided the Infection Prevention and Control (IPC) document [2], which consists of some required steps about water waste management, hand hygiene practices, safe healthcare waste management, safe management of dead bodies, and many other COVID-19 prevention requirements. Additionally, medical and health care fields are the most affected fields during COVID-19, which also consists of hospitals services and hospitals staff. However, there is a high demeaned in increasing the rate hospitals hygiene during COVID-19 pandemic to protect hospitals staff and hospital patients from COVID-19 infection. However, the availability of the personal protective equipment (PPE), such as hand sanitizer, Mask and gown, has become a worldwide concern, especially for health care workers. However, PPE can provide safe health-care services for hospital patients during COVID-19 pandemic. Moreover, WHO and all followed health ministries in whole world have increased there continues concern about offering all hygiene requirements, and tests for the hospitals to ensure of a high health safety level during the pandemic. However, the process of hospitals hygiene ensuring can take times and effort, but recently machine learning and deep learning has integrated in many COVID-19 researches to support health fields. relatively, introducing machine learning algorithms and methods in predicting the rate of hospitals hygiene can has great impact on enhancing hospitals and health fields services especially during COVID-19. The main contribution in this work is to build a machine learning model that can predict the rate of hygiene in hospitals during COVID-19, in which the most of the studies in the research area were focused on studying hand hand hygiene during COVID-19. As a sample we have used the COVID-19 Nursing Home Dataset, which consists of a USA counties hospitals monitoring information. The experiment has answered three major questions: which city has the most confirmed COVID-19 cases, which city has the most confirmed COVID-19 cases in the hospitals staff, which city has the high commitment in supply of PPE during COVID-19, is there any relation between COVID-19 cases and hygiene commitment. However, We have proposed a feature extraction, and comparing the results estimating from K-means clustering algorithm, and three classification algorithms: random forest, decision tree, and Naive Bayes, for predicting the hospital's hygiene rate during COVID-19 using COVID-19 Nursing Home Dataset. However, the results show that classification algorithms have addressed better in prediction than clustering, in which Naive Bayes considered the best algorithm for achieving the research goal. AS a result the research has discovered that the hospitals, that offered weekly amount of PPE have pass the personal quality test, which lead to decrease the number of COVID-19 cases between the hospitals staff. The work has structured as follows: Section 2 consists of the related studies, Section 3 includes the materials and methods. Section 4 is results and discussion. The last section concludes the work.

II. Related Studies

In most recent study [3] researcher conducted a review study for Coronavirus Disease-2019 (COVID-19), in his study he mentioned some recommendations, one of them is that people should be asked to practice hand hygiene frequently every 15–20 min. Hand hygiene compliance remains under the expectation and gets low among people who work in healthcare centres and visitors [4]. According to [5], statistics and recent studies indicate the rates of hand hygiene compliance in medical organisations, which hastate from 5 to 89%, with average compliance of 38.7%. The rate of Hand hygiene compliance almost stands on the method employed for measurement and place of assessment, with differentiation between clinics, intensive care units (ICU) and other places in the medical organisations. Several studies emphasis on that covering medical include its different unites by hand hygiene dispensers is not the ideal way to gain a high level of hygiene compliance [6][7]. That indicates the importance of using technologies to support decision-makers to allocate hand hygiene dispensers and get the vital data to monitor and manage hygiene systems in an optimal way [8]. [9] discuss in an explorer study several strategies and techniques for hand hygiene monitoring include direct observation, Video-assisted direct observation passive monitoring and automated individual monitoring. Furthermore, they reported and compared different hand hygiene monitoring systems in critical voice. In terms of using smart and automatic systems, several studies discuss and investigate using electronic solution for managing and monitoring hand hygiene compliance in hospitals and medical centers [10]. One of the applied solutions to improve the monitoring of hand hygiene system is an electronic device attached to a hand hygiene container to record the frequency of using that device. The device records each event by stamping time and date then transferred to a central computer to analysis and print a report related to that device. However, the electronic counting devices have a simple function to do it, which is counting events without more information like who has accessed the device, also cannot provide information about hand hygiene compliance [11]. The health sector is one of the main domains affected by those changes and takes advantages of that development in several medical fields, including sterilization within medical facilities. Much research conducted in this branch [12], [13]. For instance, [14] apply machine learning algorithm to investigate if location, time-based factors, or other behavior data can determine what characteristics are predictive of hand washing non-compliance events. Therefore, they found that the prediction of compliance can be done using those factors and would support decision-makers to make planes and decisions by using the predicted scenarios and results.

III. Materials and Methods

This section, have introduced a detailed explanation about the research methodology, dataset selection, the proposed model, data pre-processing, and programming tools. The work has been conducted on HP Pavilion laptop 8GB RAM, 1 TB HDD storage with windows 10 as an operating system. On the other hand, Anaconda with Jupiter Lab 1.2.6 has been used as the main platform for Python 3 programming language with sklearn library for K-means clustering and decision tree algorithms, and Rapid Miner 9.8 for random forest and naive Bayes algorithms. However, Rapid Miner is a data science and machine learning software that can deal with noisy data in a robotic manner. Moreover, open refine 3.4 has used for data cleaning, which explained in the next subsection as part of data pre-processing steps. The research has proposed a feature extraction and classification model for predicting the hospital's hygiene rate during COVID-19 using COVID-19 Nursing Home Dataset. However, the research has achieved by follow a well-defined methodology shown in Fig. 1, started by dataset selection, then data pre-processing, feature extraction, and attributes correlation study, after that one clustering algorithm: K-means clustering, and three classification methods: Decision Tree, Naive Bayes classifier, and Random Forest has applied, and performance has evaluated using the coefficient matrices.

A. Data

This subsection has presented information about the experiment dataset. The model has validated using the COVID-19 Nursing Home Dataset shown in Table I, which was offered by The Nursing Home COVID-19 Public File in, supported by The Centers for Medicare & Medicaid Services (CMS) and CDC's National Healthcare Safety Network (NHSN) system. The dataset consists of 87 columns with 245870 records that include different information about USA counties and provider cities, resident impact, facility capacity, staff, and supplies of protective equipment during the COVID-19 outbreak. Additionally, the first deadline for data reporting by USA counties hospital facilities was on the 17 of May 2020 at 11:59 p.m., where the dataset last updated date was the 17 of September 2020. However, data pre-processing and feature extraction have been explained in detail in the next section.

TABLE I. COVID-19 Nursing Home Dataset.

1- Dataset general information	
Total attributes	87
Total size	245870
Owner	CMS-Division of Nursing
Category	Special Programs-Initiatives
License	Public Domain U.S. Government
Contact Email	NH_COVID_Data@cms.hhs.gov
2- The experimental data	
Total attributes	12
Total size	120172
3- Extracted attributes description	
Pcity	Provider city
PQAT	Passed quality assurance test
WN95N	One week supply of N95 masks
WSM	One week supply of surgical masks
WEP	One week supply of eye protection
WGO	One week supply of gowns
WGL	One week supply of gloves
WHS	One week supply of hand sanitizer
TRC19	Total resident confirmed COVID-19
STC19	Staff total confirmed COVID-19
STD19	Staff total COVID-19 death
WTPPE	Weekly personal protection equipment

B. Data Pre-Processing and Feature Extraction

This section discussed the main steps of preparing the dataset before model implementation. Moreover, the feature extraction step, explained in this section is the main feature extraction, as a part of data processing, but in order to implement the models, the step has repeated in different ways depending on each model's requirements. Data pre-processing step has started by changing the dataset columns names in the Excel file, in which each column was referred as a

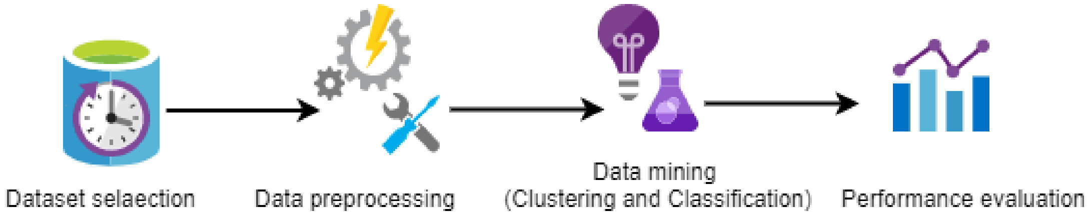

Fig. 1. Research Methodology

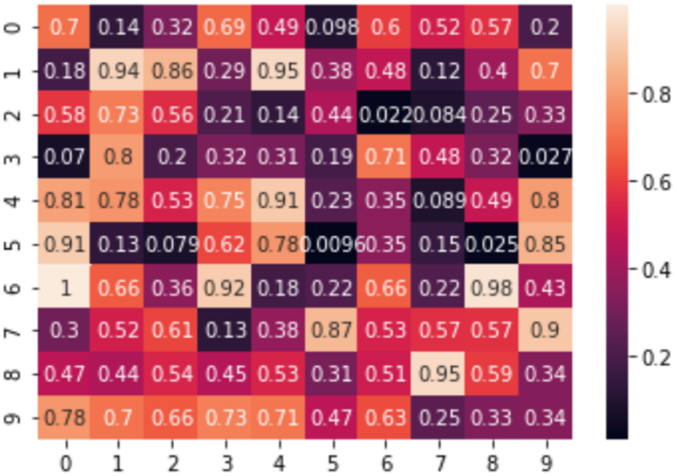

Fig. 2. The Relationship between Three Attributes: Staff Total Conferred COVID-19 Cases (STC19), Staff Total COVID-19 Death (STD19), and Total Resident Conferred COVID-19 Cases (TRC19)

combination of 3 to 4 words, the names has shorten by take the first letter in each word to simplify file reading step in the Python code Table I part 3 shows some examples of this step. The next step was eliminating the empty records and cells in the dataset, in which the COVID-19 Nursing Home Dataset was full of null values as a result from the empty records and cells. To illustrate this, open refine 3.4 has used to delete all the empty records and refill the empty cells with 'N' if the column data type is character, or '0' if the column data type is number. Additionally, during the model's implementation step, data transformation and replacement has done with the means of all other values, was the best solution. Furthermore, the outliers have eliminated from the dataset using Interquartile Range (IQR) in Python, and the outliers values have appeared in 13 columns with IQR values. Last step of data pre-processing was eliminating the duplicate values, which done on python, and decreased the total number of cleaned records to 120172 as shown in Table I. On the other hand, data reduction is one of the most important steps in any data pre-processing, in which it assists in reducing the data dimensionality, that lead to increasing model performance, identifying the irrelevant data, simplifying data visualisation, and enhancing results predicting. However, feature extraction is one of the well-known methods in data reduction, and we have applied two techniques of feature extraction depending on the dataset needs. We have started by feature selection, or it can refers as attributes selection as that the research did not follow any feature selection approaches or algorithms the step has done by selecting the target attributes from the dataset depending on dataset knowledge and the experimental requirements. The extraction was achieved using a python script in JupyterLab, and we have successfully extracted 11 attributes from the dataset as shown in Table I (3). Furthermore, feature construction has done in two steps on a specifies dataset attributes: WN95N, WSM, WEP, WGO, WGL, WHS, Firstly, we have transformed the values in each attribute to '0' for 'N', and '1' for 'Y' so it became a numerical values, then we have applied an aggregation process on the attributes, because they all representing the availability of PPE, which represented in one attribute: weekly total personal protective equipment (WTPPE), that has values in the interval of 0 to 6, shown in Table I (3). Before implementing the model a 500 objects has extracted from the dataset as a sample for studying the correlation and relationship among dataset attributes. However, multiple types of visualization have used in JupyterLab such as horizontal line charts and heatmap. To illustrate this, the horizontal line chart in Fig. 3 shows the relation between USA counties provider cities, counties hospital staff total conferred COVID-19 cases in the blue line, and counties residents total conferred COVID-19 cases in the red line. The curves implied a high addressing rate in staff cases, and residents in Shawnee counties, and a stabilized rate from Appleton to Drumright, where the curve returned to increase in Portland for staff cases. Consequently, the curves in Fig. 3 have come to an agreement, in that Shawnee counties have addressed the highest COVID-19 cases whether in hospitals staff, or counties residents, also the curves show the same stabilized rate in some cities, which implied a medium to high correlation between these attributes. Depending on the results presented in Fig. 3, a 10*10 Heatmap shown in Fig. 2 indicating a high correlation between three attributes: staff total conferred COVID-19 cases (STC19), staff total COVID-19 death (STD19), and total resident conferred COVID-19 cases (TRC19) confirmed. However, data description and correlation study step has been accomplished as part of data pre-processing, because each one of these attributes can give a specified descriptive for COVID-19 situations inside and outside the hospitals, which can be linked with the level of hospital hygiene.

C. K-Means Clustering

This section introduced the K-means clustering model as a unsupervised learning algorithm, which was implemented in our research. However, among all data mining research, clustering analysis research is the most influential one because naturally things and people are clustered in multiple classes

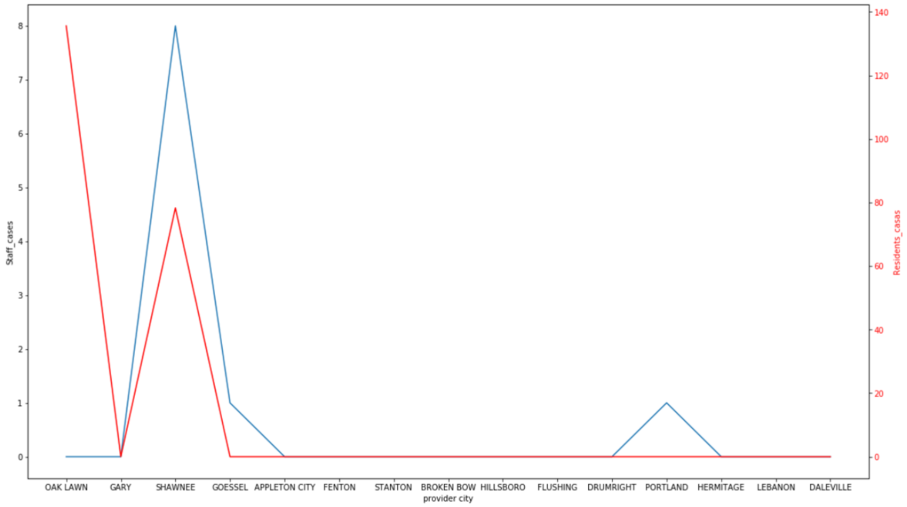

Fig. 3. The Chart Shows Three Different Attributes Provider Cities Names in the X axis, Hospital Staff Total Conferred COVID-19 Cases in the Blue Line, and Countries Residents Conferred COVID-19 Cases in the Red Line

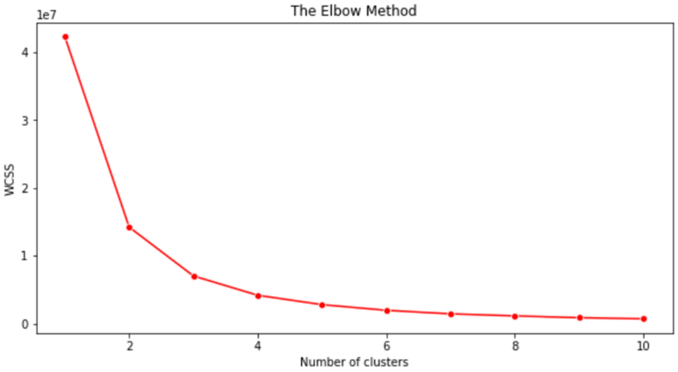

Fig. 4. The K-WCSS Curve Shows an explicit Inflection when K = 2 for the Elbow Method

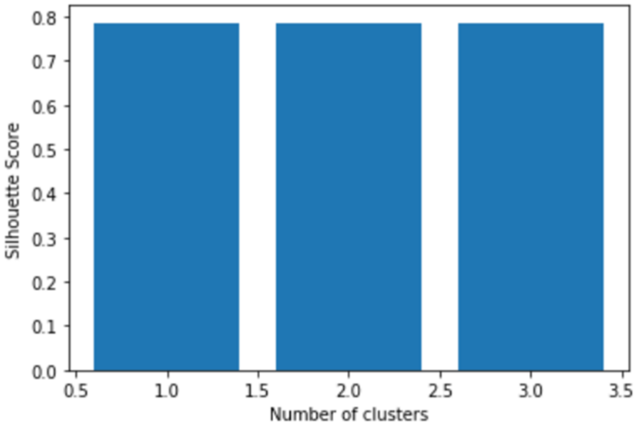

Fig. 5. The Silhouette Score Method Results (When K=1 and K=2)

and groups. Relatively, K-means clustering is one of the most used clustering methods for its fast convergence and simplicity. However, it's an iterative algorithm that uses the distance as a metric, classes of K in the data set, which represent the number of clusters, the mean of the distances, and a giving initial centroid for each class. The main problem in K-means implementation is selecting the appropriate K-value, which directly affects the algorithm result. The study [15] argued on the types of K-value selecting algorithms, such as that the Elbow , silhouette coefficient, gap statistic, and canopy. The author found that Elbow Method records the least execution time among the other methods. In this work a K-means clustering has built to cluster the hospital's hygiene depending on two attributes: PQAT, and WTPPE, that shown in Table I (3). In the first step the most relevant dataset attributes for our target attribute PQAT has examine, and that has been achieved by features selection methods. However, the "SelectKBest" method from "sklearn" library in Python has used to discovered 4 features the best relevant to PQAT: PCity, STC19, STD19, WTPPE, and WTPPE has chosen to be the second attribute in the 2D K-means model, in which the total size of records used in K-means clustering was 70673. Then as inspired from [15], we performed a comparison between two of K-values estimated methods: the elbow method, and silhouette score method. The main idea behind the elbow method is using, the square of the distance between sample points. Each cluster and its centroid gives a series of K values. The sum of squares error (SSE), or the within cluster sum of squares (WCSS) are used as a performance indicator, and Iterate the value of K to calculate the SSE or WCSS, where the smaller value of SSE or WCSS represent, that each cluster is similar. The K-WCSS curve in Fig. 4 shows the result of applying the elbow method on our model, and the curve indicated an explicit inflection in K = 2. On the other hand, the silhouette score method measures the similarity of an object to its cluster compared to the other clusters, where silhouette score high value indicates that a well matched object in the same clusters, and poorly matched with the other cluster. Fig. 5, shows the result of calculating the best number of k-values that can enhance the performance of our K-means clustering model.

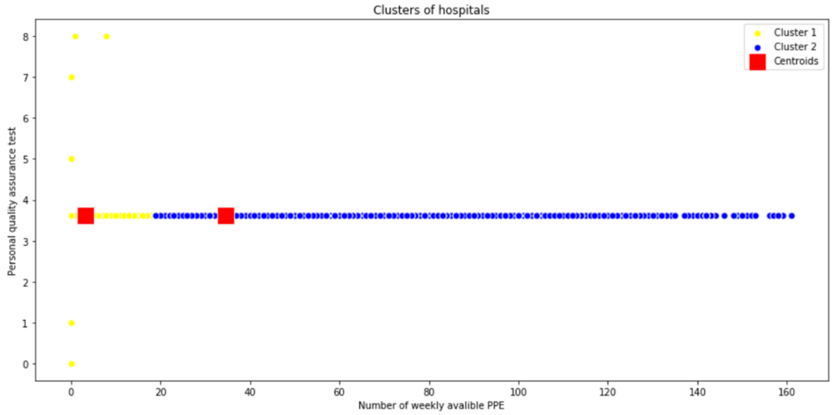

Fig. 6. The Hospitals Clusters according to Weakly available PPE, and Personal Quality Assurance Test

D. K-Means Clustering Measurement Metrics

This section explained the measurement metrics used to evaluate the K-means clustering model. This work has used six different clustering K-means clustering:

- **Adjusted rand index (ARI):** Measured the corrected version of rand index, which is a statistical measurement that represents the similarity between two data clustering. The ARI value establishes a minimum or starting point for comparisons between two pairs of clusters.
- **Mutual Information based scores (MI):** Statistically it measures the strength of association between two variables.
- **Homogeneity, completeness and v-measure (HCV):** Represent the quality of clustering algorithms.
- **Fowlkes-mallows scores (FMS):** Evaluate the similarity between clusters.
- **Calinski-harabaz index (CHI):** It is the ratio of the sum among the clusters, the value indicates better algorithm performance.
- **Davies-bouldin score (BD):** represents the ratio among cluster scatter, and cluster separation, often used to evaluate the best number of clusters, lower value indicates better clustering.

The K-means algorithm has repeated in the range of K-values from 1 to 2, and each time the algorithm measurement metrics have evaluated using the 6 previous metrics, which are shown in Table II (2). According to Table II, using 2 clusters is more suitable with our data. However, according to Fig. 6 there were two clusters for the hospitals PQAT, and WTPPE, but because of how the clusters have visualized, in which each cluster appear in one line, which emphasized that the K-means clustering is not suitable on the attributes (PQAT, and WTPPE).

E. Classification Methods

This section have introduced in detail the classification methods that have been implemented in this work. Classification is one of data mining techniques that aim to determine dataset records classes depending on the value of the predefined target attribute from the dataset. However, there are many classification methods that proved their benefits in prediction and identified various life problems. To illustrate this, Ira Ekanda Putri [16] used two different classification methods: naive Bayes classifier, support vector machine (SVM) to predicate heart disease, in which SVM was addressed the best prediction performance. On the other hand, Dedy Hartama [17] used the C4.5 decision tree to predicate patterns of interest of high school graduates. The study was conducted on a new admission student's dataset consisting of: student ID, student name, school location, and school type. The experiment was done on a rapid miner studio to build the C4.5 decision tree, which records a good performance in classifying students into their interesting study program according to the given attributes in the dataset. Moreover, unlike the supervised data mining methods such as clustering algorithms, the classification methods are supervised methods, which need to divide the dataset into train and test, and use labeled classes for model training steps. In this work three classification methods have chosen: Decision Tree, Naive Bayes classifier, and Random Forest, and the dataset has divided into 70% train data and 30% for the test data.

TABLE II. The execution time for K-value estimating methods and K-means clustering measurement metrics.

1-Methods execution time		
K-value	Elbow method	Silhouette score
1	1.6 s	-
2	2.4 s	1388.2 s
2- Measurement metrics		
K-value	Metrics	Score
1-2	ARI	0 - 0.29
1-2	MI	-1.2e-15 - 0.43
1-2	HCV	1 - 1
1-2	FM	0.6 - 0.68
1-2	CH	1.59e+06 - 1.59e+06
1-2	DB	0.5 - 0.5

*1) **Decision Tree**:* The first classification model used in this work is the decision tree, which is one of the most known classification methods. However, decision trees have

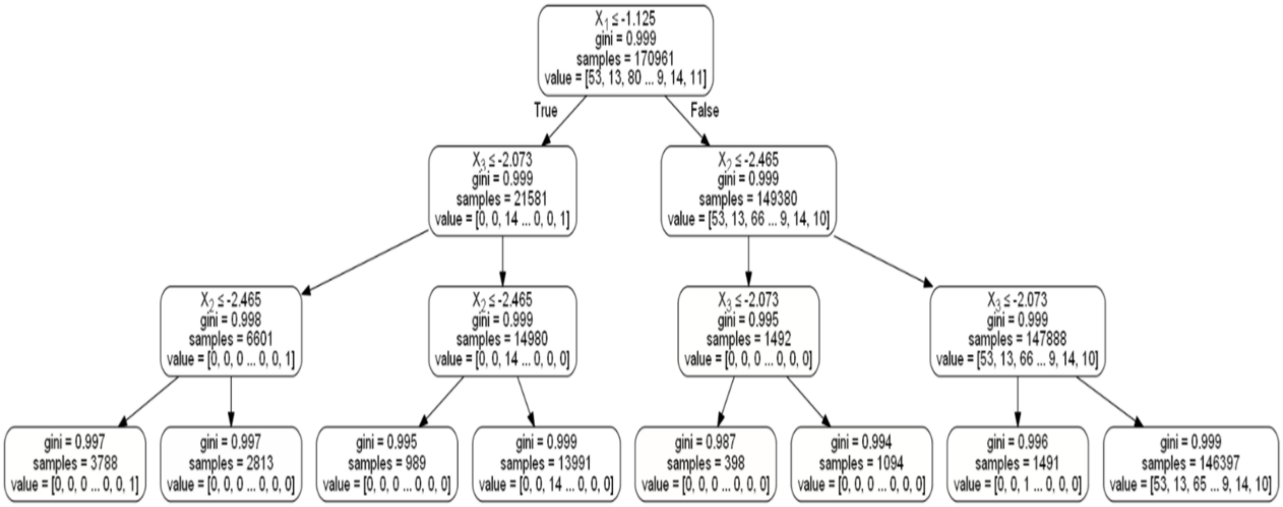

Fig. 7. Decision Tree

many advantages compared to other classification methods, in which it is inexpensive to construct, extremely fast at classification of unknown records, and easy to interpret for small dataset, but it can take time in the training step. In this work the decision tree algorithm has applied on COVID-19 Nursing Home dataset using Jupyter Lab 1.2.6 as a python 3 programming environment. Additionally, four attributes have extracted from COVID-19 Nursing Home Dataset, which are SQT, STD19, TRD19, and STC19. The algorithm starts by selecting the best split and when to stop splitting for the tree. One of the most known measurements used to measure the node impurity, and the best split of the decision tree is Gini index (Gini coefficient). Additionally, the Gini index shown in Equation 1, is a measurement, which calculates the probability rate of a particular feature, that is wrongly classified when randomly selected.

$$Giniindex = 1 - \sum_{i=1}^{n} (p_i)^2 \qquad (1)$$

However, Fig. 7 shows the resulting decision tree that was built to predict the target attribute PQT according to STD19, TRD19, and STC19. Additionally, the Gini index is estimated for each node in the tree, and in each branch the tree takes a sample from the dataset. However, the branches have been labeled into true or false. However, the tree has stopped splitting with 8 leafs which presented the final prediction results.

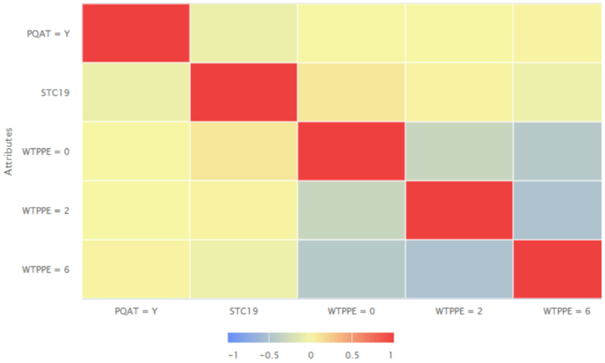

Fig. 8. Naive Bayes Classifier Correlation Matrix for Three Attributes: PQT, WTPPE, and STC19

2) ***Naive Bayes classifier:*** The second classification method, which has been used in this work is Naive Bayes classifier. Naive Bayes classifier is a straightforward classification method that does not require any complex rules for implementation. To illustrate this, Naive Bayes depend on probability theory for finding the best classification class for the target attribute. However, the main advantages of Naive Bayes, that it is a robust method for isolating noise points in the dataset, handling the missing values by ignoring the missing values during calculating the probability, and it has the ability for dealing with the irrelevant attributes. A study by Aji Prasetya Wibawa [18] has proposed a Naive Bayes model for classifying journals Quartile. The study included 1491 records and 10 attributes collected from "Journal Rankings in the Scimago Journal and Country Rank", and specified in computer science. The model was addressed with a good prediction rate with accuracy approximately equal to 71.60%, which concluded that Naive Bayes had the ability to classify journals Quartile even if that the accuracy was not optimal. On the other hand, in this work Naive Bayes has applied on the dataset using Rapid Miner 9.8, and three attributes have selected: PQT, WTPPE, and STC19, in which PQT is the target and WTPPE, STC19 are the predictor attributes. Fig. 9 shows the results of Naive Bayes classifier in predicting the PQT using WTPPE and STC19. To illustrate this, in Fig. 9(A) the result shows that the hospitals that have high availability of personal protective equipment equal to 6 have high probability to pass the personal assurance test, where hospitals that have low than 6 WTTP have high probability to fail in the PQT. On the other hand, Fig. 9(B) shows that there is increasing rate in the probability of pass the PQT when the number of conferred COVID-19 cases in the hospitals staff (STC19) is less than 25, then the probability of pass decreased while the number of STC19 increased. Additionally, in Fig. 8 we have examined

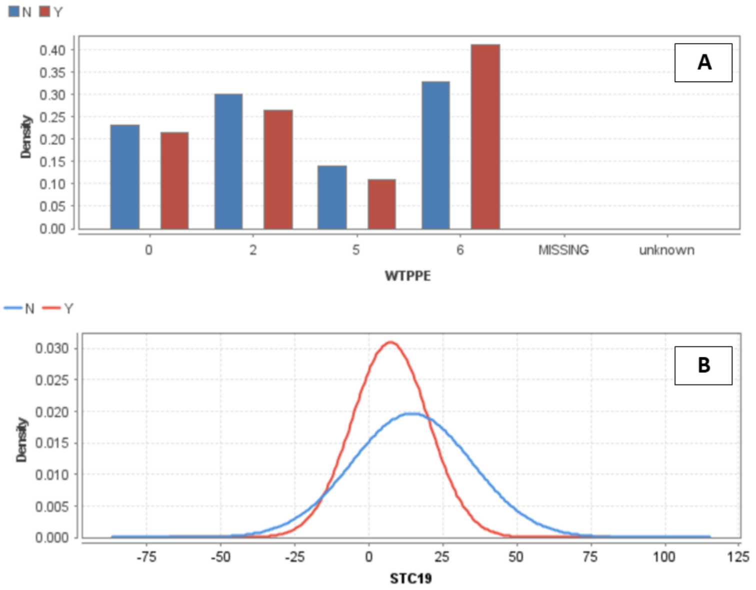

Fig. 9. Naive Bayes Classifier Result, (A) Predicting PQT Attribute depend on WTPPE Attribute, (B) Predicting PQT Attribute depend on STC19 Attribute

the correlation between the three attributes used in the Naive Bayes classifier, the heatmap shows a high correlation between PQT, WTTP, and STC19.

3) ***Random forest*** *:* The last classification model, that has applied on the dataset is random forest classifier. However, random forest is a combination of predictors trees, in which the value of each tree depends on a random vector sampled separately and with the same distribution for all trees in the random forest. The algorithm was discovered by Leo Breiman [19]. Moreover, a study by Ramón Díaz-Uriarte [20], has used a random forest algorithm to select the relevant gene expression for sample classification, such as distinguish between cancer and non-cancer patients. The author argued that random forest recorded an optimal performance in predicting and classification genes expression, and it has the ability to deal with noisy data such as the micro-array data. For this work the random forest model have applied for predicting the target attribute PQT using the STC19 attribute using Rapid Miner 9.8. However, Fig. 10 shows a snapshots of the resulted random forest, in which 20 trees have applied in the predicting model, each tree has classified the hospital's hygiene according to PQT. The figures show that the trees from 1 to 12 have predicted that most hospitals with staff confirmed COVID-19 cases have passed the personal quality test, except if the staff cases are larger from 137,500 or less than 102.500 then the hospitals have not passed the PQT. The result obtained in tree 12 in Fig. 10 has the same result obtained previously by the decision tree model in Fig. 7. However, at tree 20 in Fig. 10 all the hospitals have passed the PQT, in which all the tree labels (leafs) became equal 'Y', and that could be not a reasonable decision.

F. Classification Models Evaluation

This section have explained in detail the performance evaluation step for the classification models used in this research. The evaluation was done using four performance metrics: accuracy, F1 score, recall, and precision.

TABLE III. The values of classification models evaluation metrics

Model	Accuracy	F1 score	Recall	Precision
Decision tree	0.1	2.1	1.2	0
Naive Bayes	98.1	99	100	98.1
Random forest	98.1	0	100	98.1

- **Accuracy:** It is defined by the ratio of the correct predicted observation of the model to the total of observations.
- **Recall:** Defined as the ratio of the correct predicted of the positive observations to the total observations in the actual class.
- **Precision:** Defined as the ratio of the correct predicted of the positive observations to the total predicted positive observations.
- **F1-Score:** It presented the average of Precision, Recall and described by Equation 2

$$2 * (Recall * Precision)/(Recall + Precision) \quad (2)$$

However, Table III shows the results of the evaluation metrics on the three classification models: decision tree, Naive Bayes, and random forest. The results shows that Naive Bayes and random forest have addressed a high accuracy in prediction of hospital hygiene with values equal to 98.1 compared to decision trees. Moreover, Naive Bayes and random forest have recorded the same value for recall and precision metrics, which equals to 100 and 98.1 respectively, where decision tree has addressed the lowest values for recall and precision metrics equals to 1.2 and 0 respectively. However, according to F1 score values we have realized that Naive Bayes classifier is the convenience classification algorithm for prediction hospitals hygiene, in which that Naive Bayes has record the highest f1

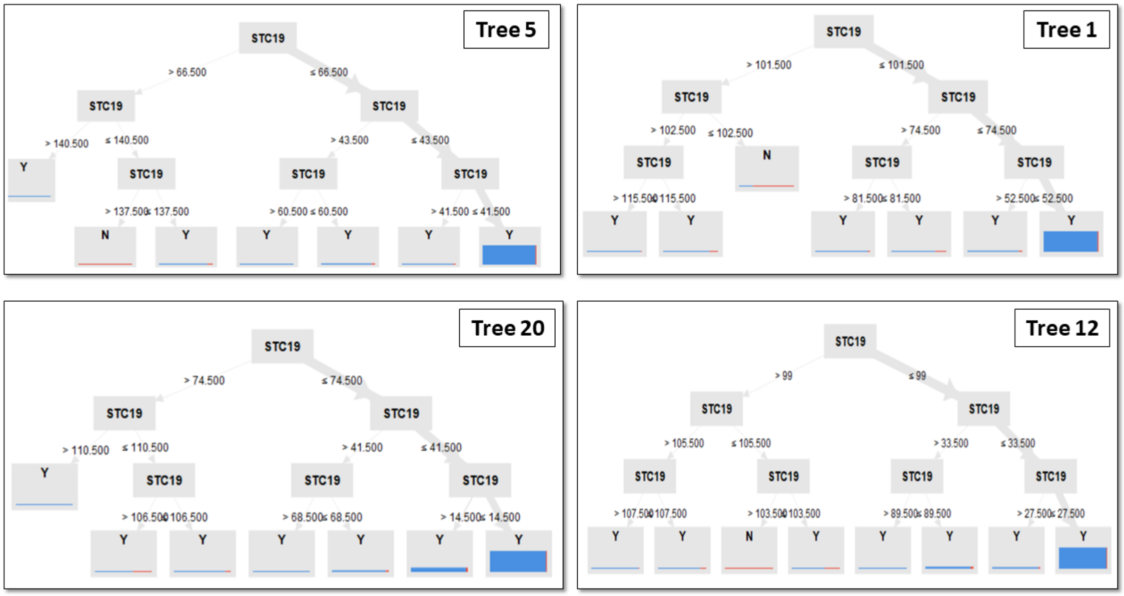

Fig. 10. Sample from Random Forest Results

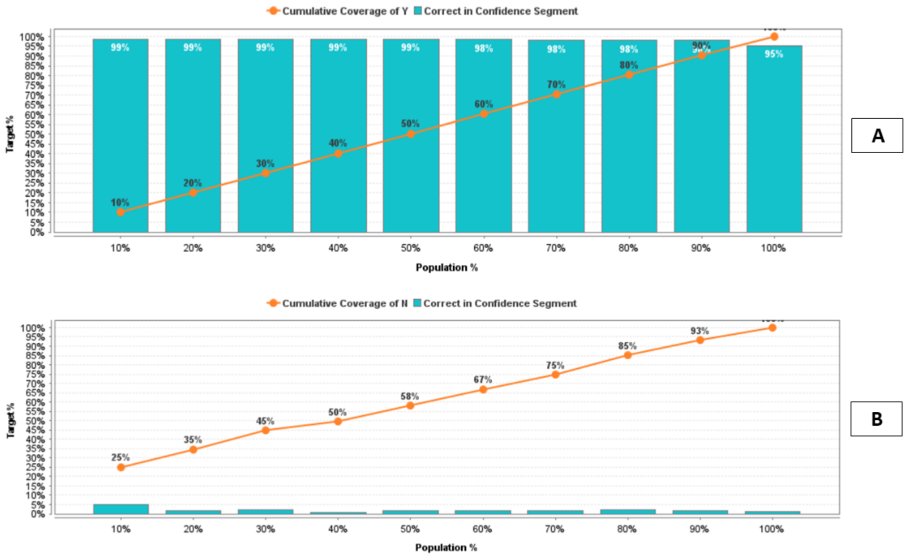

Fig. 11. Life Chart (A) Naive Bayes Classifier, (B) Random Forest

score equals to 99, where the other classification algorithms have addressed 2.1 for decision tree and 0 for random forest. On the other hand, Fig. 11 shows the life chart for both Naive Bayes and random forest, in which Naive Bayes life chart has recorded the best prediction and coverage values. To illustrate this, in Fig. 11 (A) the ratio of correct confidence and prediction with coverage have reached to the highest values with the increase in the life chart, in which the coverage has reached to 100%, and the correct confidence has reached to 95%. However, Fig. 11 (B) shows a random forest low value for the correct confidence equals to %5 corresponding to a high value for coverage equals to %100. As a result according to Table III and Fig. 11 we have considered, that Naive Bayes is the best classifier in prediction hospitals hygiene on the COVID-19 Nursing Home dataset.

IV. Results and Discussion

This section have discussed the results obtained from K-means clustering as a clustering algorithm, and Naive Bayes as classification algorithms, which has recorded the best prediction result compared to random forest and decision trees. The aim from this research is to predict the hospital's hygiene rate during COVID-19 using COVID-19 Nursing Home Dataset. According to the result from K-means clustering on hospitals PQAT, and WTPPE, which shown in Fig. 6, the research has emphasized that the K-means clustering is not suitable for the research problem. Additionally, the main limitation of K-means clustering is the long execution time required to implement the algorithm on a big dataset. On the other hand, according to Table III and Fig. 11, Naive Bayes is the best classifier in prediction hospital hygiene on the COVID19 Nursing Home dataset. As a result Naive Bayes is the best algorithm that is suitable to the research dataset, and to study the hospital's hygiene. The research has discovered that the hospitals that offered weekly amounts of PPE have passed the personal quality test, which lead to a decrease the number of COVID-19 cases between the hospital's staff.

V. Conclusion

This research, have proposed a feature extraction, and comparing the results estimating from K-means clustering algorithm, and three classification algorithms: random forest, decision tree, and naive Bayes, for predicting the hospital's hygiene rate during COVID-19 using COVID-19 Nursing Home Dataset. However, most of the studies in the research area were focused on studying hand hand hygiene during COVID-19. Additionally, the results show that classification algorithms have addressed better in prediction than clustering, in which naive Bayes considered the best algorithm for achieving the research goal with accuracy value equal to 98.1%. As a conclusion from this research hospitals that offered weekly amounts of PPE have passed the personal quality test, which lead to a decrease in the number of COVID-19 cases between the hospital's staff.

VI. Acknowledgment

The authors extend their appreciation to The Researchers Supporting Project number (1-441-68) Taif University, Taif, Saudi Arabia

References

[1] S. Chen, J. Yang, W. Yang, C. Wang, and T. Bärnighausen, "Covid-19 control in china during mass population movements at new year," *The Lancet*, vol. 395, no. 10226, pp. 764–766, 2020.

[2] W. H. Organization *et al.*, "Water, sanitation, hygiene, and waste management for sars-cov-2, the virus that causes covid-19: interim guidance, 29 july 2020," World Health Organization, Tech. Rep., 2020.

[3] Z. Y. Zu, M. D. Jiang, P. P. Xu, W. Chen, Q. Q. Ni, G. M. Lu, and L. J. Zhang, "Coronavirus disease 2019 (covid-19): a perspective from china," *Radiology*, p. 200490, 2020.

[4] S. Debnath, D. P. Barnaby, K. Coppa, A. Makhnevich, E. J. Kim, S. Chatterjee, V. Tóth, T. J. Levy, M. d Paradis, S. L. Cohen *et al.*, "Machine learning to assist clinical decision-making during the covid-19 pandemic," *Bioelectronic medicine*, vol. 6, no. 1, pp. 1–8, 2020.

[5] J. Kirk, A. Kendall, J. F. Marx, T. Pincock, E. Young, J. M. Hughes, and T. Landers, "Point of care hand hygiene—where's the rub? a survey of us and canadian health care workers' knowledge, attitudes, and practices," *American journal of infection control*, vol. 44, no. 10, pp. 1095–1101, 2016.

[6] P. Parchure, H. Joshi, K. Dharmarajan, R. Freeman, D. L. Reich, M. Mazumdar, P. Timsina, and A. Kia, "Development and validation of a machine learning-based prediction model for near-term in-hospital mortality among patients with covid-19," *BMJ Supportive & Palliative Care*, 2020.

[7] H. Burdick, C. Lam, S. Mataraso, A. Siefkas, G. Braden, R. P. Dellinger, A. McCoy, J.-L. Vincent, A. Green-Saxena, G. Barnes *et al.*, "Prediction of respiratory decompensation in covid-19 patients using machine learning: The ready trial," *Computers in biology and medicine*, vol. 124, p. 103949, 2020.

[8] L. J. Conway, "Challenges in implementing electronic hand hygiene monitoring systems," *American Journal of Infection Control*, vol. 44, no. 5, pp. e7–e12, 2016.

[9] A. Marra and M. Edmond, "New technologies to monitor healthcare worker hand hygiene," *Clinical Microbiology and Infection*, vol. 20, no. 1, pp. 29–33, 2014.

[10] J. M. Boyce, "Current issues in hand hygiene," *American journal of infection control*, vol. 47, pp. A46–A52, 2019.

[11] S. Hagel, J. Reischke, M. Kesselmeier, J. Winning, P. Gastmeier, F. M. Brunkhorst, A. Scherag, and M. W. Pletz, "Quantifying the hawthorne effect in hand hygiene compliance through comparing direct observation with automated hand hygiene monitoring," *infection control & hospital epidemiology*, vol. 36, no. 8, pp. 957–962, 2015.

[12] J. Conway, "The industrial internet of things: an evolution to a smart manufacturing enterprise," *Schneider Electric*, 2016.

[13] V. Tsiatsis, S. Karnouskos, J. Holler, D. Boyle, and C. Mulligan, *Internet of Things: technologies and applications for a new age of intelligence*. Academic Press, 2018.

[14] P. Zhang, J. White, D. Schmidt, and T. Dennis, "Applying machine learning methods to predict hand hygiene compliance characteristics," in *2017 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI)*. IEEE, 2017, pp. 353–356.

[15] C. Yuan and H. Yang, "Research on k-value selection method of k-means clustering algorithm," *J—Multidisciplinary Scientific Journal*, vol. 2, no. 2, pp. 226–235, 2019.

[16] I. E. Putri, D. Rahmawati, and Y. Azhar, "Comparison of data mining classification methods to detect heart disease," *Pilar Nusa Mandiri: Journal of Computing and Information System*, vol. 16, no. 2, pp. 213–218, 2020.

[17] D. Hartama, A. P. Windarto, and A. Wanto, "The application of data mining in determining patterns of interest of high school graduates," in *Journal of Physics: Conference Series*, vol. 1339, no. 1. IOP Publishing, 2019, p. 012042.

[18] A. P. Wibawa, A. C. Kurniawan, R. P. Adiperkasa, S. M. Putra, S. A. Kurniawan, Y. R. Nugraha *et al.*, "Naïve bayes classifier for journal quartile classification," *International Journal of Recent Contributions from Engineering, Science & IT (iJES)*, vol. 7, no. 2, pp. 91–99, 2019.

[19] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[20] R. Díaz-Uriarte and S. A. De Andres, "Gene selection and classification of microarray data using random forest," *BMC bioinformatics*, vol. 7, no. 1, p. 3, 2006.