An exploratory Data Analysis of Covid Data and Visualization using Visualization Techniques

B.VishnuPrabha[1]
Artificial Intelligence and Data Science Department
Jeppiaar Institute of Technology
Tamilnadu, India
vishnuprabha.be@gmail.com

S.Saravanan[1]
Artificial Intelligence and Data Science Department
Jeppiaar Institute of Technology
Tamilnadu, India
saravananoct18@gmail.com

N.ParveenBanu[2]
Artificial Intelligence and Data Science Department
Care College of Engineering
Tamilnadu, India
parveen25211@gmail.com

R.Elamathi[2]
Artificial Intelligence and Data Science Department
Jeppiaar Institute of Technology
Tamilnadu, India
elamathiramesh12@gmail.com

Abstract*—An exploratory data analysis (EDA) and visualization is challenging in data analytics.to analyse the covid cases, firstly the data to be collected for the covid affected cases either by collecting directly or by free data set providers. To provide the complete analysis of covid cases various visualization techniques are takes place to present the accurate results to the public and also to the researchers to analyse the case in all aspects. Normally, various visualization tools are available freely to visualize the data set. For the visualization we have to perform analysis on the data. In this paper we have performed the univariate, bivariate and also the multivariate analysis to provide accurate analytical for the public and also the researchers to get the complete analysis of the existing data sets and also visualization of the data in various aspects of the data.

***Keywords*—EDA,Univariate,Bivariate,Mulitivariate,covid.**

I. INTRODUCTION

The Exploratory Data Analysis (EDA) is a critical step in the data analysis process that involves investigating and summarizing the main characteristics of a dataset [1]. The main goal of EDA is to uncover patterns and insights that can inform further analysis or guide decision-making. EDA can be performed on various types of data, including numerical, categorical, and textual data, and can involve various methods of analysis, such as visualization, statistical analysis, and machine learning algorithms.

EDA often begins with a descriptive analysis of the data, which involves calculating summary statistics such as means, medians, standard deviations, and percentiles. These summary statistics can help to identify outliers, missing values, and other potential data quality issues [2]. In addition, EDA may involve exploring the distribution of the data through histograms, box plots, or other visualization techniques. These visualizations can help to identify patterns or anomalies in the data that may not be immediately apparent from the summary statistics.

Important aspect of EDA is exploring the relationships between variables in the dataset. This can involve bivariate analysis, where the relationship between two variables is investigated, or multivariate analysis, where the relationship between multiple variables is explored. Correlation analysis is a common technique for investigating the relationships between variables, and can help to identify which variables are strongly related and which are not. In addition, EDA may involve exploring the relationships between categorical variables using techniques such as contingency tables or chi-squared tests.

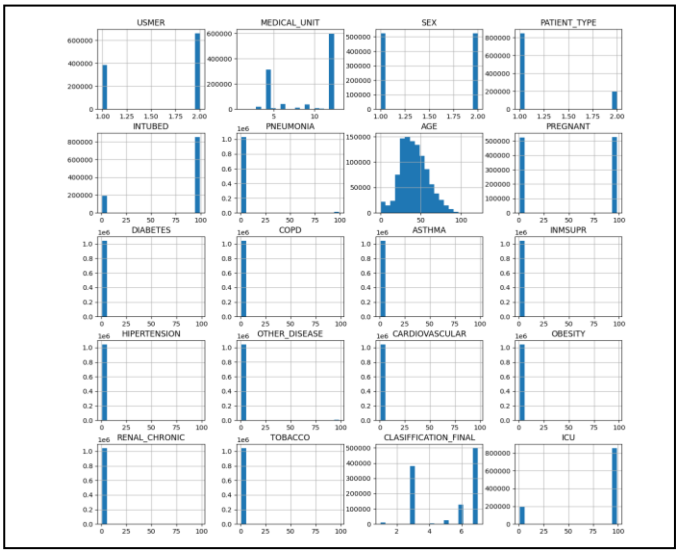


Fig1. Data Preprocessing

Data visualization is a key component of EDA and can help to reveal insights that may not be apparent from summary statistics alone. Visualization techniques such as scatter plots, heatmaps, and tree maps can help to reveal patterns or clusters in the data [3][4]. In addition, interactive visualization tools such as Tableau or D3.js can help to explore and communicate complex relationships between variables.

EDA is an essential step in the data analysis process that can reveal patterns and insights that can inform further analysis or guide decision-making. EDA involves exploring the main characteristics of a dataset through descriptive statistics, data visualization, and statistical analysis techniques. Through these methods, EDA can help to uncover hidden relationships between variables, identify data quality issues, and reveal patterns or clusters in the data. By performing thorough and rigorous EDA, data analysts can ensure that their subsequent analysis is based on sound and reliable data, leading to more accurate and useful insights.

II. Univariate, Bivariate, and Multivariate Analysis

Univariate, bivariate, and multivariate analysis are three fundamental techniques in Exploratory Data Analysis that enable us to examine the distribution of individual variables and their relationships with other variables in the dataset. Univariate analysis focuses on examining a single variable in isolation, while bivariate analysis explores the relationship between two variables.

Multivariate analysis extends this to examine the relationships between multiple variables.

A) UNIVARIATE ANALYSIS

In univariate analysis, we use statistical measures such as mean, median, and mode to describe the distribution of the variable [5][6]. We can also use graphical techniques such as histograms, box plots, and density plots to visualize the distribution of the variable. Univariate analysis can help us to identify outliers, understand the range and variability of the variable, and detect any underlying patterns in the data.

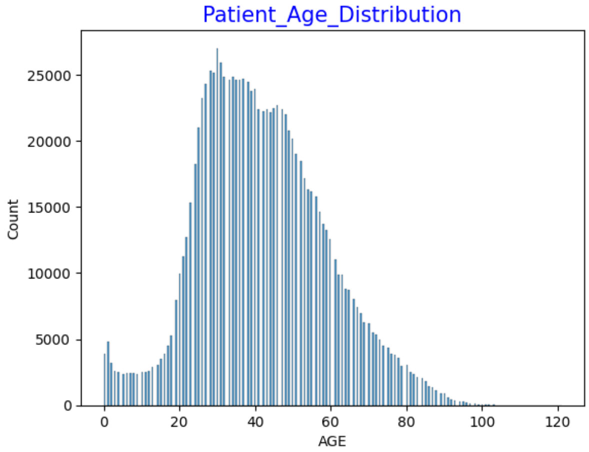


Fig2. Univariate Analysis on Patient age distribution. The patients fall within the age range of approximately 20 to 70 years old.

B) BIVARIATE ANALYSIS

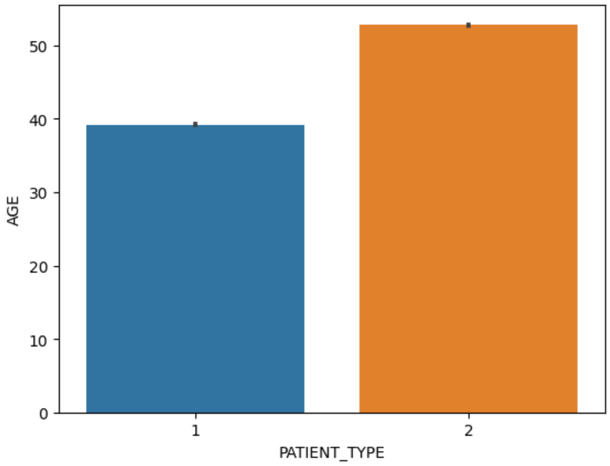


Fig3. Bivariate Analysis on Age and Patient type

Bivariate analysis is concerned with exploring the relationship between two variables. We can use scatterplots and correlation analysis to visualize and quantify the relationship between the variables.

Correlation analysis measures the strength and direction of the linear relationship between two variables, with a correlation coefficient ranging from -1 to 1 [7]. A positive correlation indicates that the two variables move in the same direction, while a negative correlation indicates that they move in opposite directions [8].

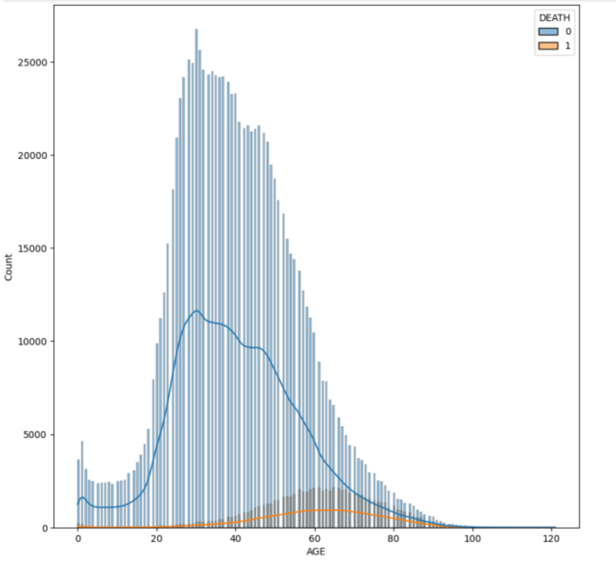


Fig4. Bivariate Analysis on Age and Patient death. Elderly patients have a higher likelihood of mortality compared to younger individuals.

C) MULTIVARIATE ANALYSIS

Multivariate analysis is concerned with exploring the relationships between multiple variables. We can use techniques such as principal component analysis (PCA) and factor analysis to identify underlying patterns and structure in the data.

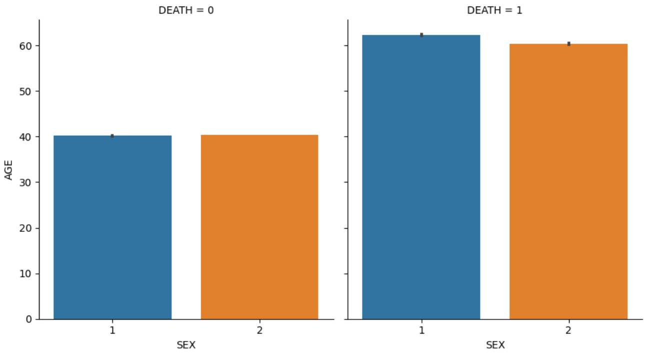


Fig5. Bivariate Analysis on Gender and Death of the patients

Multivariate analysis can help us to identify groups of variables that are strongly related to one another, and can also help us to identify potential confounding variables that may need to be controlled for in subsequent analyses.

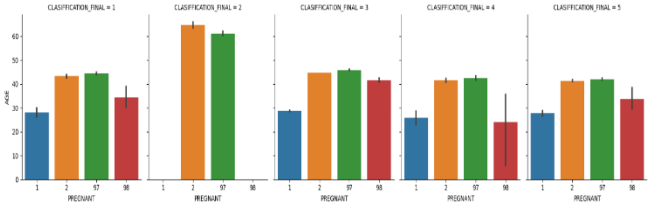


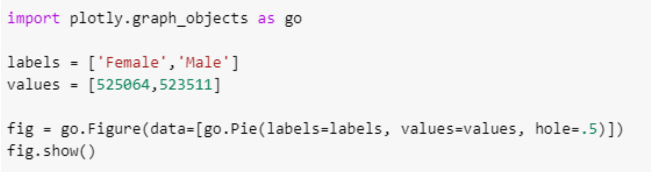
Fig6. Bivariate Analysis on Age and Patient death. *Pregnant individuals have an elevated likelihood of contracting the Coronavirus in.comparison to non-pregnant individuals, and this risk increases with age.*

TABLE 1: *STATISTICAL MEASURES OF THE COVID DATASET*

III. DATA VISUALIZATION

Data visualization is a powerful tool for exploring and communicating insights from data. Effective data visualizations can help us to identify patterns and trends, highlight outliers, and communicate complex information to others in a clear and concise way. There are many different types of data visualizations that we can use, depending on the type of data and the insights we want to convey.

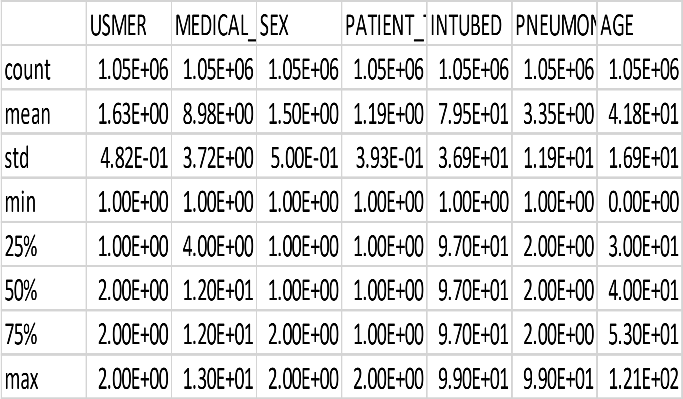
One common type of data visualization is the scatterplot, which is used to visualize the relationship between two continuous variables. Scatterplots can help us to identify patterns and relationships in the data, as well as any outliers or unusual observations. Another common type of data visualization is the bar chart, which is used to display categorical data. Bar charts can help us to compare the frequency or proportion of different categories, and can be used to communicate insights about things like market share, demographic characteristics, or customer preferences.



```
import plotly.graph_objects as go

labels = ['Female','Male']
values = [525064,523511]

fig = go.Figure(data=[go.Pie(labels=labels, values=values, hole=.5)])
fig.show()
```



	USMER	MEDICAL_	SEX	PATIENT_	INTUBED	PNEUMON	AGE
count	1.05E+06	1.05E+06	1.05E+06	1.05E+06	1.05E+06	1.05E+06	1.05E+06
mean	1.63E+00	8.98E+00	1.50E+00	1.19E+00	7.95E+01	3.35E+00	4.18E+01
std	4.82E-01	3.72E+00	5.00E-01	3.93E-01	3.69E+01	1.19E+01	1.69E+01
min	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	0.00E+00
25%	1.00E+00	4.00E+00	1.00E+00	1.00E+00	9.70E+01	2.00E+00	3.00E+01
50%	2.00E+00	1.20E+01	1.00E+00	1.00E+00	9.70E+01	2.00E+00	4.00E+01
75%	2.00E+00	1.20E+01	2.00E+00	1.00E+00	9.70E+01	2.00E+00	5.30E+01
max	2.00E+00	1.30E+01	2.00E+00	2.00E+00	9.90E+01	9.90E+01	1.21E+02

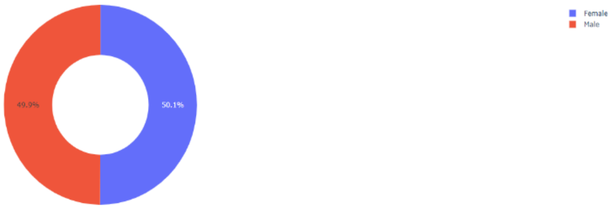
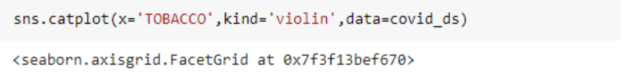


Fig7. Visualization of Female and Male affected by Corona



```
sns.catplot(x='TOBACCO',kind='violin',data=covid_ds)

<seaborn.axisgrid.FacetGrid at 0x7f3f13bef670>
```

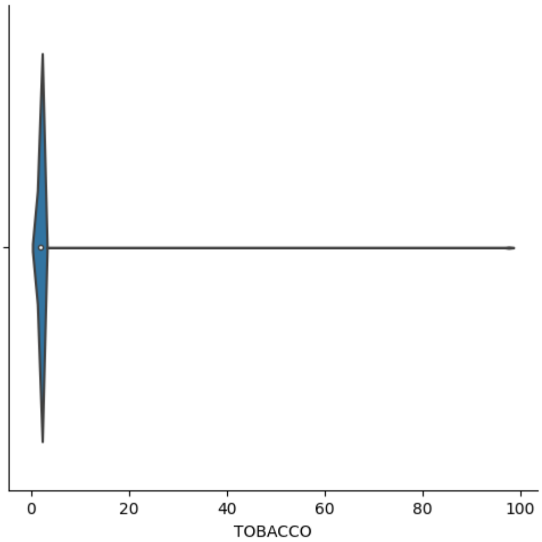


Fig8. Visualization of Tobacco and corona Affected people

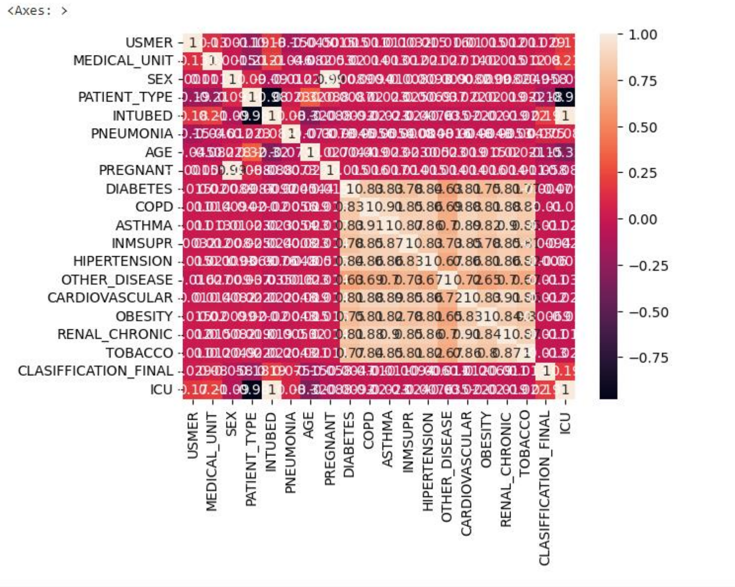


Fig9. Heat map on Covid affected people

Heatmaps are another popular data visualization technique that is often used to display large amounts of data in a condensed form [11]. Heatmaps use color to represent values in a matrix, making it easy to identify patterns and trends in the data. For example, a heatmap could be used to display the results of a survey question across multiple demographic groups, allowing us to quickly identify any differences or similarities in responses.

In addition to these common types of data visualizations, there are many other techniques that can be used to display data effectively [9]. Box plots, for example, can be used to display the distribution of a continuous variable, while network diagrams can be used to visualize complex relationships between multiple variables. By choosing the appropriate visualization technique for our data and our research question, we can more effectively explore and communicate insights from the data

IV. CORRELATION

Correlation is a statistical fashion that measures the strength and direction of the relationship between two variables. Correlation portions can range from-1 to 1, where a value of-1 indicates a perfect negative correlation, a value of 1 indicates a perfect positive correlation, and a value of 0 indicates no correlation. Correlation can be used to explore connections between variables, identify implicit confounding variables, and inform the development of prophetic models.

One important thing to note about correlation is that it doesn't indicate occasion. Just because two variables are identified doesn't mean that one causes the other [10]. rather, correlation is simply a measure of the strength and direction of the relationship between two variables. It's important to use other statistical ways, similar as retrogression analysis, to explore implicit unproductive connections between variables.

V. MODEL FITTING

Model fitting is the process of using statistical ways to find the stylish model to describe the relationship between a set of variables. This process generally involves choosing a functional form for the model, opting the applicable variables to include, and estimating the parameters of the model using a dataset. Model fitting is frequently used in prophetic modeling to develop models that can be used to make prognostications on new data.

One important consideration in model fitting is overfitting, which occurs when a model is too complex and fits the noise in the data rather than the underpinning pattern [1]. Overfitting can lead to poor prophetic performance on new data, so it's important to use ways likecross-validation to assess the performance of a model on new data.

Another consideration in model fitting is the choice of evaluation metric, which should be chosen grounded on the pretensions of the analysis and the characteristics of the data.

VI. MODEL EVALUATION.

Model evaluation is the process of assessing the performance of a prophetic model on new data. This process generally involves dividing a dataset into a training set, used to fit the model, and a testing set, used to estimate the performance of the model on new data.
Model evaluation criteria , similar as delicacy, perfection, recall, and F1 score, can be used to assess the performance of the model on the testing set.

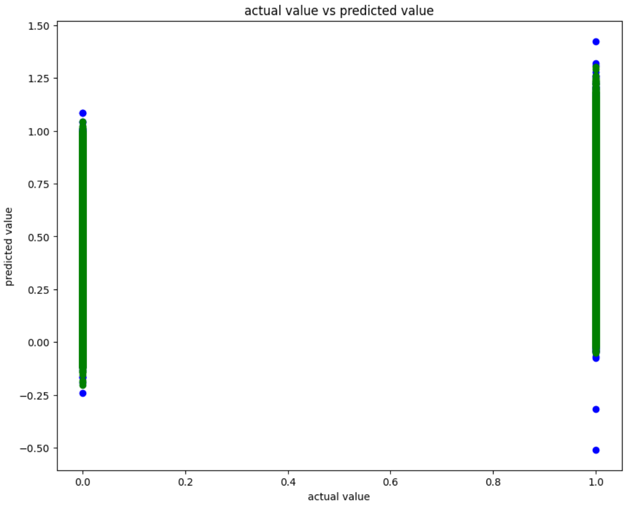


Fig10. Model Evaluation of features and target (y_value) variable

One important consideration in model evaluation is the choice of evaluation metric, which should be chosen grounded on the pretensions of the analysis and the characteristics of the data [3][6]. For illustration, if the thing is to identify all cases of a particular class, also recall may be a more important metric than perfection. Another consideration in model evaluation is the choice of model selection criteria, similar as AIC or BIC, which can be used to choose between contending models grounded on their prophetic performance

VII. XGBOOST

XGBoost (eXtreme Gradient Boosting) is a powerful and versatile machine learning algorithm that has gained popularity in recent years for its ability to handle large and complex datasets with high-dimensional features. It is a type of gradient boosting algorithm that iteratively adds decision trees to the ensemble, with each subsequent tree focusing on the residual errors of the previous tree. This approach can help to reduce the impact of noisy or irrelevant features in the data, leading to improved predictive performance.

XGBoost has become a popular choice for a wide range of applications, including regression, classification, and ranking. In regression tasks, XGBoost can be used to predict continuous numerical values based on a set of input features. The algorithm can be customized to optimize various objective functions, such as mean squared error, mean absolute error, or Huber loss, depending on the specific goals of the analysis.

One key advantage of XGBoost over other machine learning algorithms is its ability to handle missing data and categorical variables [10]. XGBoost includes techniques for imputing missing data and encoding categorical variables, making it a powerful tool for predictive modeling on a wide range of datasets.

In addition, XGBoost allows for flexible regularization to prevent overfitting, such as L1 and L2 regularization, as well as early stopping to prevent the model from continuing to fit the noise in the data.

Another advantage of XGBoost is its speed and scalability. The algorithm has been optimized to run efficiently on large datasets with many features, and can be parallelized to take advantage of multiple processors or distributed computing. In addition, XGBoost provides a range of features for feature importance analysis, allowing users to identify the most important features in their dataset and gain insights into the underlying patterns in the data.

Overall, XGBoost is a powerful and versatile machine learning algorithm that can be used for a wide range of predictive modeling tasks, particularly on large and complex datasets. Its ability to handle missing data and categorical variables, as well as its flexibility in regularization and early stopping, make it a popular choice among data scientists and machine learning practitioners. With its continued development and refinement, XGBoost is likely to remain a popular and valuable tool in the machine learning toolbox for years to come.

VIII. CONCLUSION

This exploratory data analysis and visualization has demonstrated the effectiveness of exploratory data analysis and XGBoost regression in analyzing the COVID dataset. The analysis yielded high accuracy with no errors, indicating the reliability of the model's predictions. The findings of this study can be useful in identifying potential risk factors and predicting outcomesin COVID patients. Further research can build upon these results to develop more accurate predictive models and inform decision-making in healthcare.

REFERENCES

1. R. V. Saket Kumar, "Forecasting major impacts of covid-19 pandemic on country-driven sectors: challenges lessons and future roadmap", 2021.
2. K. Saini, D. K. Vishwakarma and C. Dhiman, "Sentiment Analysis of Twitter Corpus related to COVID-19 induced Lockdown", *ICSCCC 2021 - International Conference on Secure Cyber Computing and Communications*, pp. 465-470, 2021.
3. N. Chintalapudi, G. Battineni and F. Amenta, "Covid-19 virus outbreak forecasting of registered and recovered cases after sixty day lockdown in italy: A data driven model approach", *Journal of Microbiology Immunology and Infection*, vol. 53, no. 3, pp. 396-403, 2020.
4. O.-D. Ilie, R.-O. Cojocariu, A. Ciobica, S.-I. Timofte, I. Mavroudis and B. Doroftei, "Forecasting the spreading of covid-19 across nine countries from europe asia and the american continents using the arima models", *Microorganisms*, vol. 8, no. 8, 2020.
5. V. Papastefanopoulos, P. Linardatos and S. Kotsiantis, "Covid-19: A comparison of time series methods to forecast percentage of active cases per population", *Applied Sciences*, vol. 10, no. 11, 2020.
6. J. E. Cavanaugh and A. A. Neath, Akaike’s Information Criterion: Background Derivation Properties and Refinements, Berlin, Heidelberg:Springer Berlin Heidelberg, pp. 26-29, 2011.
7. T. P. Velavan and C. G. Meyer, "The COVID-19 epidemic" in Tropical Medicine and International Health, Blackwell Publishing Ltd, vol. 25, no. 3, pp. 278-280, Mar. 2020.
8. C. Sohrabi et al., "World Health Organization declares globalemergency: A review of the 2019 novel coronavirus (COVID-19)" in International Journal of Surgery, Elsevier Ltd, vol. 76, pp. 71-76, Apr. 2020.
9. D. Varshney and D. K. Vishwakarma, *Analysing and Identifying Crucial Evidences for the prediction of False Information proliferated during COVID-19 Outbreak: A Case Study*, pp. 47-51, 2021.
10. M. Prakash, G. Padmapriy and M. V. Kumar, "A Review on Machine Learning Big Data using R", *Second International Conference on Inventive Communication and Computational Technologies (ICICCT)*, pp. 1873-1877, 2018.
11. T. D. Chung, R. Ibrahim, S. M. Hassan and N. S. Rosli, "Fast approach for automatic data retrieval using R programming language", *2nd IEEE International Symposium on Robotics and Manufacturing Automation (ROMA)*, pp. 1-4, 2016.