Machine Learning Approach to Predict the Trends of the COVID-19 Pandemic: A survey

Utkaarsh Bhaskarwar[1], Manasi Variar[2], Ashwith Poojary[3]

[1-3]BE Information Technology Student - Pillai College of Engineering, New Panvel, Maharashtra

---***---

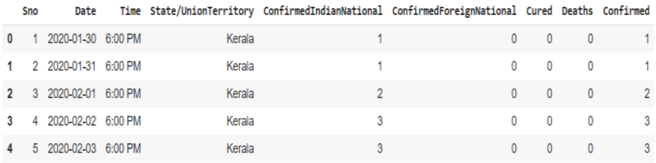
Abstract – *The COVID-19 pandemic has led to a dramatic loss of human life worldwide and presents an unprecedented challenge to public health, food systems, and therefore the world of labor. India had its first case of Covid 19 on 30th January 2020. Cases of the COVID-19 pandemic are exponentially increasing day by day within the whole world. As of June 2020, India has the 2nd highest number of confirmed cases in the world. It has become important to reduce the number of cases and save as many lives as possible. Therefore, if the number of deaths is predicted early, Millions of lives could be saved. Government can predict the spread of infections, resulting in better planning of resources, better preparedness for response, and improved health care facilities. Machine learning plays a very important role in pandemic situations such as predicting the future death toll, cured cases, and thereby planning further measures based on these predictions. Many models of machine learning have been proposed by various authors in the literature. This motivated us to present a survey of those models and implement them for comparison in this work.*

***Keywords* -- Prediction, Machine Learning, COVID-19, Regression, Neural Networks.**

1.INTRODUCTION

COVID-19 is a disease caused by a new coronavirus called SARS-CoV-2. WHO first became aware of the new virus on 31st December 2019, after receiving a report of a case group of "viral pneumonia" from Wuhan, China. As of May 8th, 2020, in India, 56,340 positive cases have been reported. India, with a population of over 1.34 billion—the second largest population within the world—will have difficulty in controlling the transmission of severe acute respiratory syndrome coronavirus 2 among its population. Multiple strategies would be highly necessary to handle the present outbreak; these include computational modeling, statistical tools, and quantitative analyses to regulate the spread also because of the rapid development of a new treatment. The Ministry of Health and Family Welfare of India has raised awareness about the recent outbreak and has taken necessary actions to regulate the spread of COVID-19. The central and state governments are taking several measures and formulating several wartime protocols to attain this goal. Moreover, the Indian government implemented a 55-days lockdown throughout the country that started on March 25th, 2020, to decrease the transmission of the virus. This outbreak is inextricably linked to the economy of the nation. After all, it has dramatically impeded industrial sectors because people worldwide are currently cautious about engaging in business in the affected regions.

The dataset [1] consists of features of COVID-19 data. It contains 15807 samples of COVID-19 cases in India from 31 January 2020 to 8 June 2021



	Sno	Date	Time	State/UnionTerritory	ConfirmedIndianNational	ConfirmedForeignNational	Cured	Deaths	Confirmed
0	1	2020-01-30	6:00 PM	Kerala	1	0	0	0	1
1	2	2020-01-31	6:00 PM	Kerala	1	0	0	0	1
2	3	2020-02-01	6:00 PM	Kerala	2	0	0	0	2
3	4	2020-02-02	6:00 PM	Kerala	3	0	0	0	3
4	5	2020-02-03	6:00 PM	Kerala	3	0	0	0	3

Fig -1. Dataset

1.1 Machine Learning

Machine learning is a branch of artificial intelligence (AI) that focuses on creating applications that learn from data and improve their accuracy over time, without the need for programming. In data science, an algorithm is a series of statistical processing steps. In machine learning, algorithms are "trained" to find patterns and features in massive amounts of data to make decisions and predictions based on new data. The better the algorithm, the more accurate the decisions and predictions, because the more data it processes.

Today, machine learning examples are everywhere. The digital assistant will search the internet and play music based on our voice commands. The website recommends products, movies, and songs based on the content we have previously purchased, viewed, or listened to. The robot vacuums while we clean the floor. Something better with our time. The spam detector prevents unwanted emails from reaching our inbox. Medical imaging systems can help doctors find tumours they may have missed. The first autonomous car is on the road.

You can expect more. As big data continues to grow and computing becomes more powerful and affordable, and as data scientists continue to develop more capable algorithms, machine learning is increasingly making our personal and professional lives more efficient.

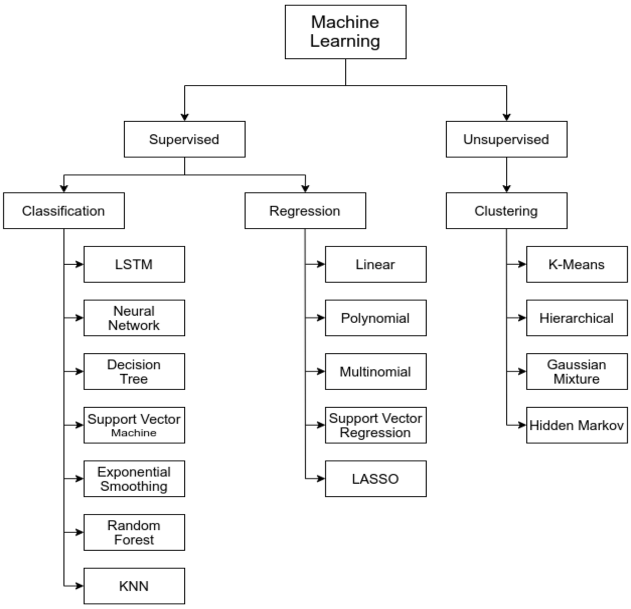


Fig -2. Machine Learning Algorithms

1.2 Regression

A regression problem is when the output variable may be a real or continuous value, like “height” or “age”. There are various regression models but the simplest one is linear regression. It tries to fit data with the simplest and best hyper-plane which matches through the points. It is used when the dataset contains only numerical variables, which means there are no categorical variables.

2. ALGORITHMS USED

Linear Regression: It is the commonly used algorithm and can be imported from the linear regression class. A single input variable is used to predict one or more output variables, provided that the input variable is not correlated together. It is expressed as: $y = m * z + b$
where y is a dependent variable, z is an independent variable, m is the slope of the best fit line that will give accurate output, and b is its intercept.

Polynomial Regression. If the data points do not fit in a linear regression model, it becomes an ideal case for polynomial regression. It is a regression algorithm that shows the relationship between self-reliant and dependent variables as nth degree polynomials. It can be stated as the modified linear regression model to increase the accuracy of the model.

Lasso regression: Lasso (Least Absolute Shrinkage and Selection Operator) is a regularization technique. It is a type of linear regression that uses reduction. Reduction is where the data values are reduced towards a central point, like the mean. Lasso uses the L1 regularization technique. Regularization is a technique that modifies so that the model generalizes better. L1 Regularization adds a penalty equivalent to the absolute value of the magnitude of the coefficients. This improves the accuracy by reducing the overfitting of the model.

Ridge regression: Ridge regression performs L2 regularization, i.e., it adds a penalty equivalent to the square of the magnitude of coefficients. It reduces the size of the values during optimization and reduces overfitting. It also has a better performance than the simple regression model.

Neural Network: Neural Network or Convolution Neural Network (CNN) is a supervised type of machine learning model. It is generally used for 2D data but can also be modified to work with 1D data. They function very much similar to the human brain. It comprises several different types of layers. First is the convolutional layer, which slides a filter over the data. This extracts the majority of the detail from the data. It is then followed by a Pooling layer, whose primary aim is to reduce the size of the convolved features. Next comes a few sets of fully connected hidden layers. This consists of weights, biases, and neurons. To avoid overfitting of the model, the data is then passed through a dropout layer which randomly drops out a few nodes from the neural network during the training process. Last is the activation function which decides whether the data should be passed to the next neuron or not.

3. LITERATURE SURVEY

Saud Shaikh et al. [6] presented a paper describing the use of various regression models to predict the spread of the COVID-19 in India. Linear and Polynomial models having different degrees were plotted for prediction on the test dataset and it was seen that the polynomial model with degree 5 tends to adapt better towards the actual values. It was observed that the higher the value of a degree, the higher was the accuracy and the lesser was the Mean Square Error. Hence, polynomial regression models with higher degrees outperform other degree models and linear models. The performance evaluation is done using Mean Absolute Performance Error (MAPE) and R-squared (R^2). They utilized Heatmaps for visualizing the correlation between the entities. For forecasting the data, they made use of Tableau. The results obtained were found to be satisfactory and could be overcome by increasing the size of the dataset and using better algorithms.

Ajinkya Kunjir et al. [7] in their paper analyzed the WHO dataset to predict the trend of the widespread COVID-19. They made use of algorithms such as Long Short-Term Memory (LSTM**)**, Convolutional Neural Network (CNN), and Decision Tree. The dataset consisted of the data of 91 days. The dataset was stripped down and only India, China, and Canada were studied. Each algorithm was tested individually for each country and was compared to the actual factors (confirmed cases, deaths, recovered). The testing was done for an unknown time series of 25 days. It was seen that the CNN model fits best and tends to overlap with the data curve of the actual dataset. Based on the R^2 score, CNN performed the best followed by the Decision Tree and LSTM at the last, for all three countries. The model performed well and provided accurate results. More countries and states along with a few more algorithms can be added for better comparative study. The models can be evaluated on a few more performance metrics such as MSE, Variance.

The paper presented by Ashish Mandayam et al. [9] focuses on a Machine Learning approach to study the spread of the COVID-19. Support Vector Regression (SVR) and Linear Regression were used. The dataset is split into multiple test-train data to observe the training rate of the model with different quantities of the training set. The evaluation parameters which were considered are R-squared (R^2), Mean Absolute Error (MAE), and Mean Squared Error (MSE); but the main concentration was on the R^2 score. For both the models, it was observed that with an increasing set of training data, the performance also increased. The SVR model was always above the original curve, and would hence predict incorrectly. The difference between the actual value and the predicted value was significantly large. The SVR model was not so reliable. The Linear model tended to coincide with the original curve and also its R^2 score was very much close to 1. Hence, the Linear model performed very well. SVR can't handle large datasets and therefore its performance was not so good.

In this paper, the authors Furqan Rustam et al. [11] performed an analysis to forecast the spread of the virus for the next 10 days using various supervised machine learning models. Linear Regression (LR), LASSO Regression, Support Vector Machine (SVM), and Exponential Smoothing (ES) are the models which were tested. To evaluate the performance of the models, they made use of R-squared (R2), Adjusted R-Squared, Mean square error (MSE), and Root mean square error (RMSE). They used the dataset from the official GitHub repository provided by John Hopkins University. They performed the prediction on each factor separately for each model. For the case of death rate forecasting, based on the R^2 score, ES performed the best followed by the LR and LASSO performing equally well, and SVM performing the worst. Even for forecasting the recovery and confirmed cases, ES led the table while the R^2 score of the SVM remained the least. Hence, it was seen that ES provided promising results and performed very well followed by the LASSO and LR providing decent accuracy, and SVM performing the worst because of a few ups and downs in the dataset values.

The paper presented by Vishan Kumar Gupta et al. [3] has performed radical analysis on detecting no. of infected cases, no. of deaths caused by Covid19. The dataset includes 2342 samples of Covid-19 cases, it included various classes therefore the performance was done using multi-class classification. Classification algorithms such as Random Forest, Support Vector Machine, Decision Tree, Multinomial Logistic Regression, and Neural networks were used. The target classes were as follows- 1) Confirmed cases 2) Death cases 3) Cured cases. The performance evaluation was done using K-fold cross-validation. The Random Forest model has outperformed the other models whereas logistic regression performance was not so accurate.

In paper [4], the authors Saksham Gera et al. have conducted a research on the forecasting of Covid19 trends using various machine learning algorithms. The dataset used in this study was 292 days. Many algorithms like Linear regression, K Nearest Neighbor, Support Vector Machine, Random Forest, Elastic Smoothing were used for the study. The paper also evaluates the performance of all the algorithms used using Root Mean Square Error(RMSE), R squared score(R^2 score), Mean Absolute Error (MAE). The result was analyzed on two different factors.1) Prediction of newly infected cases 2) Prediction of death cases. For the prediction of new cases as well as deaths ES performance was better than all the other algorithms whereas SVM performed poorly. The reason for SVM performing poorly was due to the presence of outliers in the dataset and ups and downs in the data set values. The performance could have been improved if the outliers would have been managed properly in the data cleaning process.

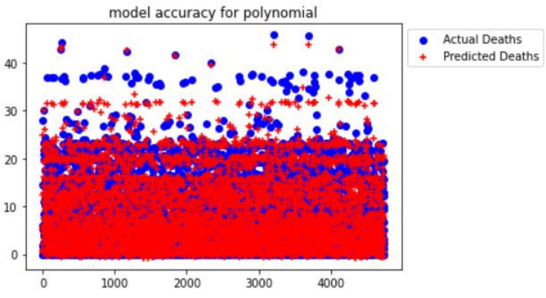
Manpinder Singh and Saiba Dalmia [2] presented a paper in which they proposed a Machine Learning model for predicting future deaths in India. The data was compiled from "Ourworldindata.org" through web scraping due to which the data is updated with time. Also, an API was created through which the number of deaths was predicted given the number of confirmed cases. The prediction was done with the help of Linear Regression and Polynomial Regression Models. Though it was noticed after the implementation that the Polynomial regression model performed exceptionally well in the performance evaluation which was done by calculating the Root Mean Square Error (RMSE) and R2 score. Since the nature of the dataset was nonlinear and the loss function and error rate was high, therefore the accuracy of the Linear Regression model was less compared to the Polynomial model.

H Zakiyyah and S Suyanto [5] in their paper have developed Machine learning models to predict the newly infected cases of Covid 19 and deaths in Indonesia. The authors have trained their models on Covid-19 Indonesia Time Series All Dataset (CITSAD) of ten provinces. Three prediction models were made using Gaussian Naïve Bayes(GNB), Support Vector Machine(SVM), and Decision Tree(DT) algorithms. The performance evaluation was done based on accuracy and processing time. Based on R2 score performance, the Decision Tree (DT) model proved to be the best model for the prediction as it had the highest accuracy and also the least processing time out of all the models being tested. The reason behind the success of the Decision tree (DT) was found out that the model could split the complex decision-making process into simpler ones making it more effective than other algorithms for more accurate prediction.

The summary of the performance evaluation metrics of all above-mentioned literature is given in Table 1.

4. METHODOLOGY

This research aims to examine the regression and to forecast the outbreak of Covid19. We obtained the dataset from Kaggle [1]. We have developed 5 machine learning models namely linear regression model, lasso regression model, ridge regression model, polynomial regression model, and neural network model. We also performed data preprocessing, transformations on the dataset. The prediction has been done based on target variables which include confirmed, cured, and death cases. Accuracy was measured using R^2 score and the error was calculated using mean absolute error (MAE). It was found that the polynomial regression model and neural network model performed better among other models. The result analysis is summarized in Table 2.



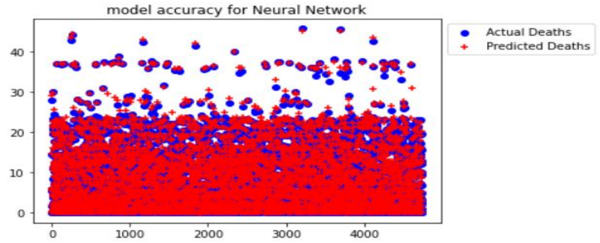


Fig - 3. Comparison for Death Cases

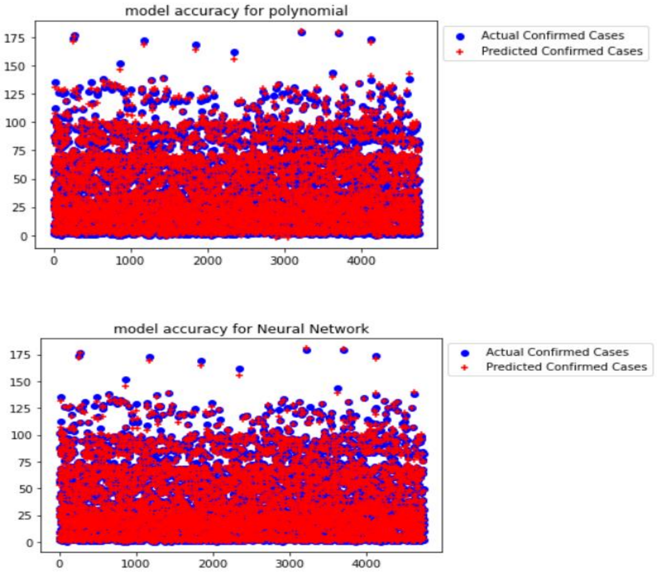


Fig - 4. Comparison for Confirmed Cases

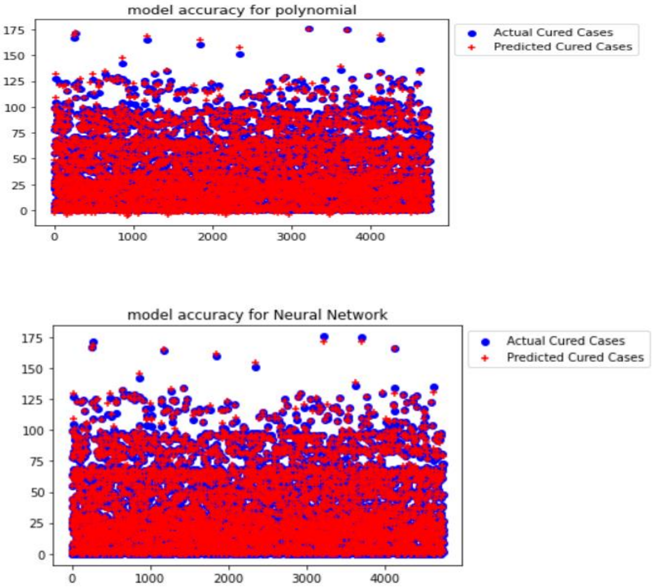


Fig - 5. Comparison for Cured Cases

5. CONCLUSION

We have surveyed various works proposed in the literature and have found that five machine learning models based on the R^2 score performed better than others. We implemented these models for the prediction of three targets viz. confirmed cases, cured patients, and deaths caused by Covid-19 in India. It was found that the Polynomial Regression model had the best accuracy of 99.7% for prediction of confirmed cases and cured cases, also, the Neural Networks model showed the least Mean Absolute Error.

Table 1. SUMMARY OF RELATED WORK.

AUTHOR NAME	ALGORITHMS USED	PERFORMANCE EVALUATION METRICS				
		R^2 score	Mean Square Error	Root Mean Square Error	Mean Absolute Error	Mean Absolute Percentage Error
Saksham Gera et al. [4]	Exponential Smoothing K Nearest Neighbor Random Forest Linear regression Support Vector Machine	0.99 0.98 0.98 0.97 0.53	11840240.11 35245068.71 13244066.11 28318061.31 5160162107.71	3440.96 5936.75 3639.23 5321.47 71834.26	17425.11 18241.53 21043.29 21437.74 55491.67	-
Ajinkya Kunjir et al. [7]	Convolutional Neural Network Long Short-Term Memory Decision Tree	0.99 0.99 -0.82	-	-	-	-
Manpinder Singh, Saiba Dalmia [2]	Linear Regression Polynomial Regression	0.99 0.99	-	0.000124 0.000010	-	-
Saud Shaikh et al. [6]	Polynomial Regression (Degree 5) Linear Regression	0.97 0.79	-	-	-	125.66 16.53
H Zakiyyah and S Suyanto [5]	Decision Tree Support Vector Machine Gaussian Naive Bayes	0.93 0.80 0.70	-	-	-	-
Ashish Mandayam et al. [9]	Linear Regression Support Vector Regression	0.98 0.80	36048939271 1.3928E+12	-	161481.82 885613.19	-
Vishan Kumar Gupta et al. [3]	Random Forest Decision Tree Support Vector Machine Neural Network Multinomial Logistic regression	0.83 0.77 0.71 0.70 0.67	-	-	-	-

Furqan Rustam et al. [11]	Exponential Smoothing	0.98	283201302.2	16828.58	8867.43	-
	LASSO Regression	0.98	234489560.99	15322.11	11693.97	
	Linear Regression	0.83	1472986504.96	38390.51	30279.55	
	Support Vector Machine	0.59	5760890969.30	75911.28	60177.90	

Table 2. RESULT ANALYSIS

Algorithms used	Confirmed		Deaths		Cured	
	R^2 score	Mean Absolute Error	R^2 score	Mean Absolute Error	R^2 score	Mean Absolute Error
Linear Regression	0.944	1.712	0.910	1.881	0.950	1.716
Lasso Regression	0.950	1.656	0.926	1.821	0.985	1.418
Ridge Regression	0.961	1.502	0.929	1.790	0.950	1.717
Polynomial Regression	0.997	1.155	0.936	1.397	0.997	1.231
Neural Network	-	0.779	-	0.423	-	0.747

REFERENCES

[1]. **Dataset Link**

[2]. M. Singh and S. Dalmia, "Prediction of number of fatalities due to Covid-19 using Machine Learning," 2020 IEEE 17th India Council International Conference (INDICON), 2020.

[3]. V. K. Gupta, A. Gupta, D. Kumar, and A. Sardana, "Prediction of COVID-19 confirmed, death, and cured cases in India using random forest model," in Big Data Mining and Analytics, 2021.

[4]. S. Gera, M. Mridul and K. Joshi, "Regression Analysis And Future Forecasting Of COVID-19 Using Machine Learnings Algorithm," 2021 11th International Conference on Cloud Computing, Data Science & Engineering (Confluence), 2021.

[5]. H Zakiyyah and S Suyanto, "Prediction of Covid-19 Infection in Indonesia Using Machine Learning Methods", 2020.

[6]. S. Shaikh, J. Gala, A. Jain, S. Advani, S. Jaidhara and M. Roja Edinburgh, "Analysis and Prediction of COVID-19 using Regression Models and Time Series Forecasting," 2021 11th International Conference on Cloud Computing, Data Science & Engineering (Confluence), 2021.

[7]. A. Kunjir, D. Joshi, R. Chadha, T. Wadiwala and V. Trikha, "A Comparative Study of Predictive Machine Learning Algorithms for COVID-19 Trends and Analysis," 2020 IEEE International Conference on Systems, Man, and Cybernetics (SMC), 2020.

[8]. E. Gambhir, R. Jain, A. Gupta, and U. Tomer, "Regression Analysis of COVID-19 using Machine Learning Algorithms," 2020 International Conference on Smart Electronics and Communication (ICOSEC), 2020

[9]. A. U. Mandayam, R. A.C, S. Siddesha, and S. K. Niranjan, "Prediction of Covid-19 pandemic based on Regression," 2020 Fifth International Conference on Research in Computational Intelligence and Communication Networks (ICRCICN), 2020.

[10]. Z. Yang and K. Chen, "Machine Learning Methods on COVID-19 Situation Prediction," 2020 International Conference on Artificial Intelligence and Computer Engineering (ICAICE), 2020

[11]. F. Rustam et al., "COVID-19 Future Forecasting Using Supervised Machine Learning Models," in IEEE Access, 2020.

[12]. KB Prakash, SS Imambi and M Ismail, "Analysis, prediction and evaluation of covid-19 datasets using machine learning algorithms", 2020 International Journal of Emerging Trends in Engineering Research , 2020.

[13]. Z. Li, S. Yang and J. Wu, "The Prediction of the Spread of COVID-19 using Regression Models," 2020 International Conference on Public Health and Data Science (ICPHDS), 2020.

[14]. K. Prathyusha, K. Helini, C. V. Raghavendran and N. Kumar Kurumeti, "COVID-19 in India: Lockdown analysis and future predictions using Regression models," 2021 11th International Conference on Cloud Computing, Data Science & Engineering (Confluence), 2021.

[15]. L. Yan et al., "Prediction of criticality in patients with severe Covid-19 infection using three clinical features: a machine learning-based prognostic model with clinical data in Wuhan", 2020.

[16]. K. Bhanu et al., "Analysis, Prediction, and Evaluation of COVID-19 Datasets using Machine Learning Algorithms," Int. J. Emerg. Trends Eng. Res, 2020.

[17]. C. Iwendi et al., "COVID-19 patient health prediction using boosted random forest algorithm," Front. Public Heal, 2020.

[18]. F. Shahid, A. Zameer, and M. Muneeb, "Predictions for COVID-19 with deep learning models of LSTM, GRU and Bi-LSTM," Chaos, Solitons and Fractals, 2020.

[19]. D. P. Kavadi, R. Patan, M. Ramachandran, and A. H. Gandomi, "Partial derivative Nonlinear Global Pandemic Machine Learning prediction of COVID 19," Chaos, Solitons and Fractals, 2020,

[20]. J. Brownlee, "Train-Test Split for Evaluating Machine Learning Algorithms." 2020.

BIOGRAPHIES

Utkaarsh Bhaskarwar – Final Year Student at Pillai College of Engineering pursuing Information Technology.

Manasi Variar - Final Year Student at Pillai College of Engineering pursuing Information Technology.

Ashwith Poojary – Final Year Student at Pillai College of Engineering pursuing Information Technology.