Forecasting Insurance and Patient Charges using Linear Regression

Jayakrishna Natarajan, Krishanu Das, Ronak Harish Patil, R B Sarooraj

Abstract: ***The insurance companies around the world work with very simple formula and have a very specific agenda. They convince people to deposit money on their name to the insurance company, in return the people are promised to be given a large sum of amount when they get an expensive hospital bill or when they meet with an accident. This amount to be paid, is generally taken from people on a monthly basis. Customers are convinced to join such a scheme as it is very tempting and the prospect of money troubles taken care of for nothing in a time of crisis seems wonderful. Insurance companies on the other hand pray that nothing happens to the customers or their families, so that they don't come looking for compensation. The money that they collect from new insurance holders is what they use to pay of the losses. Data analysis is the process of understanding the behaviour of a certain dataset when measured against certain static quantities. In this paper we are proposing to use Data science and in particular regression analysis, to analyse a dataset of patients and devise a method to predict their insurance amount. There are various types of learning and broadly speaking linear regression comes under supervised learning. We have a dataset consisting of over 1300 patients each with 7 characteristics like smoker or not, do they have children, their age, sex, BMI, etc. We are also proposing to devise methods to overcome the shortcomings of Linear regression like multicollinearity and homoscedascity, and perform the required data cleaning..***

Keywords: Data Science, Linear Regression, Learning, Analysis, Attributes

I. INTRODUCTION

Insurance is a term used when one party (the customer), pays on a monthly/yearly/one time basis, a certain amount of money to another party (insurance company), and in return this another party, will pay them a certain pre-defined large sum of money when anything happens to the commodity in discussion. In layman terms, let's say that a customer goes to a car insurance company and pays some x number of rupees to them and gets a car insurance. This money may have to be paid on a monthly or yearly basis. After some amount of time, could be years or months, if the car that customer owns, meets with an accident, the damage repair cost, will be entirely or partially covered by the insurance company, and the customer will not have to pay the same from his/her own pocket. This example can be applied to a house insurance company, life insurance, any other property insurance company and so on. Each one of the above works in their own way, and customer gets benefits in their own way, but the core of the process remains the same. Both the parties in discussion have a lot to gain from this. Customer obviously has something to gain from the fact that he may not have to pay a large sum of money all of a sudden, but will get money from the company when an accident happens. The company's gain is much more indirect. The assumption that the company makes is that, not all products get damaged, not everyone meets with an accident, and so on. So, these companies make profits if most of the customers never get an opportunity to claim the insurance. In this paper we will dealing with life insurance for patients. Life Insurance deals with people paying money to insurance companies, in the hope that if they ever die of an accident, or some natural causes before a certain age, then their families will be compensated for their sudden demise by a large sum of money. Life Insurance is generally opted by working class people where the annual family income is generated by one person, and the whole family depends on it. The amount life insurance company will pay your beneficiaries after your death depends on various factors. These factors include medical history, driving records, smoker or not, past records, etc. based on these factors, a fixed amount is set, which will be paid to your beneficiaries when you die. The dataset available has over 1300 patients, with 7 attribute values each. These attributes tell information about the patient, if he/she has children, smokes, their age, gender, BMI, etc. We will be using Linear Regression[4] analysis to analyse the available dataset and forecast the patient charges for the insurance company. The shortcomings of linear regression[4] like multicollinearity[6], homoscedascity[5], normality and other assumptions will be first taken care of by doing data cleaning on the dataset, and then the analyses will be done.

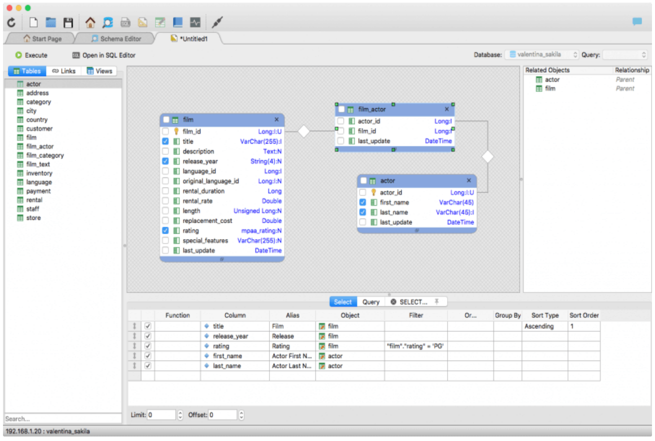
II. RELATED WORK

Nowadays, with the improving technology and resources available, the development of Machine Learning algorithms, Artificial Intelligence, Deep Learning with Neural networks, almost all of the industries have started using these novel techniques in their processes. The insurance companies are no different, and in fact these are one of the main industries where further advancements using machine learning and data analysis[2] techniques could be made.

Revised Manuscript Received on March 30, 2020.
***** Correspondence Author

Jayakrishna Natarajan, Department of Computer Science & Engineering from SRM Institute of Science & Technology, Chennai, India.
Krishanu Das, Department of Computer Science & Engineering from SRM Institute of Science & Technology, Chennai, India.
Ronak Harish Patil, Department of Computer Science & Engineering from SRM Institute of Science & Technology, Chennai, India.
Mr. R B Sarooraj, Assistant Professor, Department of Computer Science and Engineering, SRM Institute of Science and Technology, Chennai, India.

Initially, the data was stored in excel sheets and there was limited scope for any intensive analysis on them. The shortcomings of such a primitive technology, brought about the rise of databases and SQL/PLSQL queries. But even SQL queries were limited by their time complexities, and the fact that they worked on the principle of rule-based[1] operations. Rule based[1] operations work on the premise of rules, and apply those rules to the database, and the queries are returned with their outputs. But with the advent of Internet of Things in the technological world, there was large amount of data that got collected but there was no way available to analyse them, and on time. And then there was the rise of Data Analysis[2], R programming and SAS[3] (Statistical Analysis Software). Constant improvements made to them allowed for the analysis of Big Data within seconds, with better efficiency and accuracy.



Function	Column	Alias	Object	Filter	Or...	Group By	Sort Type	Sort Order
	title	Film	film				Ascending	1
	release_year	Release	film					
	rating	Rating	film	"film"."rating" = 'PG'				
	first_name	Actor First N...	actor					
	last_name	Actor Last N...	actor					

Fig 1: SQL

Like all technology though, there was one major drawback which was not yet solved. Even they were rule-based[1] and there was no scope for Unsupervised or Supervised learning on their part. But with machine learning algorithms, we can make the machine analyse the dataset, understand its behaviours and predict the future results based on its learning. The insurance companies rely on these algorithms to predict the correct amount that they can set as insurance coverage, so that they can make the most profit. There is already a lot of related work in this field, but this paper focuses on the linear regression[4] analysis, and proposes to improve the existing work and make a more efficient analysis.

III. CONTRIBUTION

This paper provides an efficient method to predict the patient charges for insurance purposes of any patient by using the data set available. With over 1300 patients and 7 attributes for each patient, the data set is quite vast and provides a deep insight to anyone interested on what factors and how determine the patient charges. With such Big Dataset, understanding them for any non-professional data scientist is very difficult. The paper provides a method to train the available dataset, and then come to a conclusion regarding patient charges. These factors of the patients can then be checked for new customers who are applying for life insurance. All the different types of analyses have some sort of limitations and assumptions made for them to work. Linear Regression[4] has few of them as well. They include homoscedascity[5], multicollinearity[6], goodness of fit[7], normality of dataset[8], etc. There are certain plots or graphs available that tell the programmer about the goodness of fit[7] and normality of dataset[8]. They are scatter plot[9], pp plot, qq plot or histogram over normal curve. RStudio or Python can be used for most of the analysis done. Lots of libraries are available in Python for the specific purpose of analysing data, as well as for other specific functionalities. The code can be written in the code shell editor, or in something called as the jupyter notebook. Python jupyter notebook (ipynb), is a very efficient IDE for writing, compiling and running the python code.

The problem of multicollinearity[6] is also easily solved by viewing the Beta-coefficients, r-squared and adjusted r-squared values are found between the variables themselves and the multicollinearity[6] is checked between the variables. If its value is below 3, there is no need to do data cleaning.

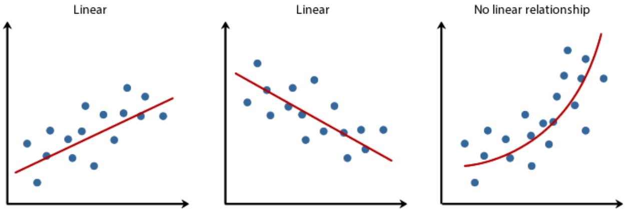
Select items to perform actions on them.

Name	Last Modified
3D Objects	2 months ago
Anaconda3	12 days ago
Cisco Packet Tracer 6.2sv	9 months ago
Contacts	2 months ago
Desktop	8 minutes ago
Documents	2 months ago
Downloads	an hour ago
Dropbox	3 years ago
env	12 days ago

Fig 2: Jupyter Notebook

Python has a lot of packages like pandas, NumPy, scikit, plotly, SciPy, statsmodels, etc. These packages are specifically present for data analysis[2], regression analysis, etc. Python also has the added advantage of ease of use. Learning python and understanding its commands are very easy for beginners and makes the analysis easily presentable to any potential client. This python analysis provides a way for insurance companies to protect themselves from any fraud and helps them predict the correct patient charges that could avoid them losses.

IV. MATHEMATICAL MODEL



Copyright 2014. Laerd Statistics.

Fig 3: X-Y Plot

The X-Y plot is the main type of graph used to depict the regression analysis pp plot.The standard straight-line graph has the same equation

$$Y = mX + C$$

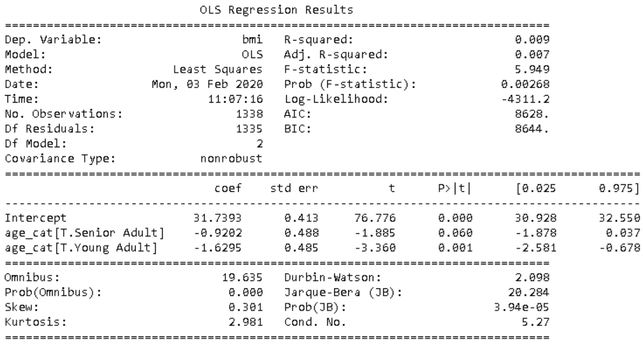
For any graph there must be a dependent variable and an independent variable. Similarly, for a pp plot, one of the axes must be independent variable.

Any non-linear relationship cannot be described by linear regression[4].

V. IMPLEMENTATION

The paper has been written on the basis of the results got from the code executed in Python 3.7. The dataset used is patient charges and forecasting them for insurance purposes.

It consists of over 1300 patients and 7 attributes for each one of them. These attributes are the patient's information like age, gender, smoker or not, medical history, if the patient has children or not, etc. First of all, the OLS (Ordinary Least Squared) results are generated. These results contain information about the f-statistic, r-squared, adjusted r-squared values, beta-coefficients, p test, degrees of freedom, etc.



```
                            OLS Regression Results
==============================================================================
Dep. Variable:                    bmi   R-squared:                       0.009
Model:                            OLS   Adj. R-squared:                  0.007
Method:                 Least Squares   F-statistic:                     5.949
Date:                Mon, 03 Feb 2020   Prob (F-statistic):            0.00268
Time:                        11:07:16   Log-Likelihood:                -4311.2
No. Observations:                1338   AIC:                             8628.
Df Residuals:                    1335   BIC:                             8644.
Df Model:                           2
Covariance Type:            nonrobust
==========================================================================================
                             coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------------------
Intercept                 31.7393      0.413     76.776      0.000      30.928      32.550
age_cat[T.Senior Adult]   -0.9202      0.488     -1.885      0.060      -1.878       0.037
age_cat[T.Young Adult]    -1.6295      0.485     -3.360      0.001      -2.581      -0.678
==============================================================================
Omnibus:                       19.635   Durbin-Watson:                   2.098
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               20.284
Skew:                           0.301   Prob(JB):                     3.94e-05
Kurtosis:                       2.981   Cond. No.                         5.27
==============================================================================
```

Fig 4: OLS Results

After viewing the OLS results, we can view the various relationships that can be generated.

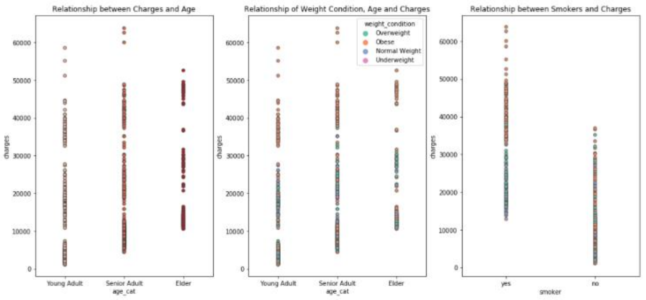
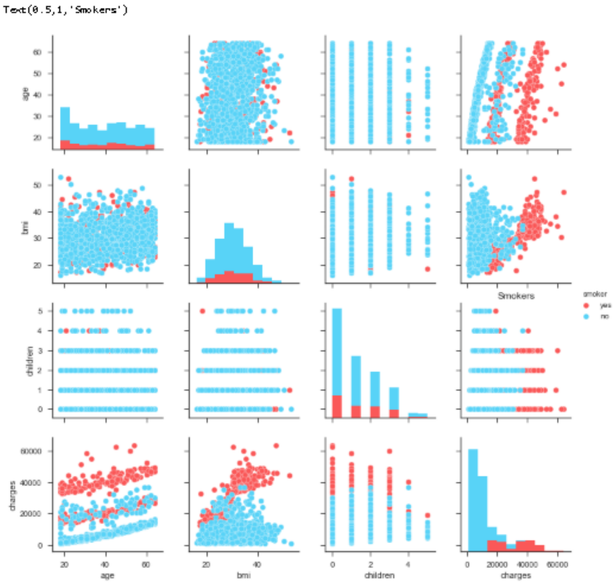


Fig 5: Relationships

These relationships describe the connection between smokers and their charges, the relationship between ages and charges and so on. Effectively these illustrations tell the user how one attribute varies due to another. Pair plot is another similar illustration.



```
Text(0.5,1,'Smokers')
```

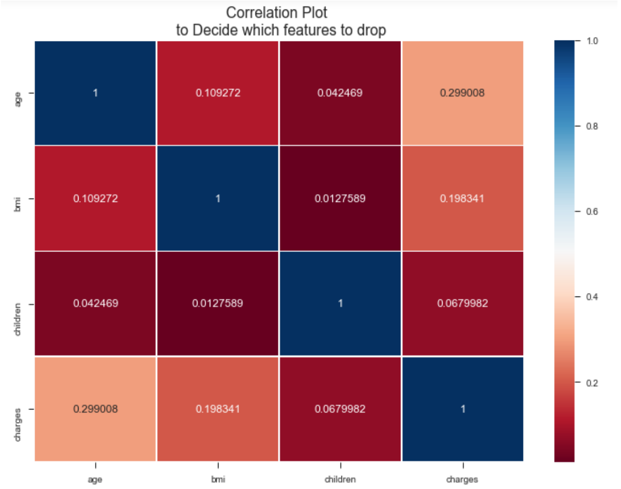
Fig 6: Pair Plot

After all these analyses to check the dataset, the next step is to remove any multicollinearity[6]. Multicollinearity[6] is the term used when there exists an interdependency between one or more independent variables with each other. This multicollinearity[6] can extremely influence the output and can cause catastrophically consequences. Multicollinearity[6] values are advised to be below 3.

Homoscedascity[5] similarly, refers to the noise or variance in the dataset. It is advisable to have constant noise throughout the data.

The correlation plot and the fitted values tell us the correlation level, which in turn can help us decide which features to drop. The various packages available in python for such analyses are plotly, SciPy, scikit, etc.

The dataset can be either under-fit, over-fit or normally fit. The goodness of fit[7] can be determined by the number of outliers and their impact on the eventual result obtained from the dataset.



	age	bmi	children	charges
age	1	0.109272	0.042469	0.299008
bmi	0.109272	1	0.0127589	0.198341
children	0.042469	0.0127589	1	0.0679982
charges	0.299008	0.198341	0.0679982	1

Fig 7: Correlation Plot

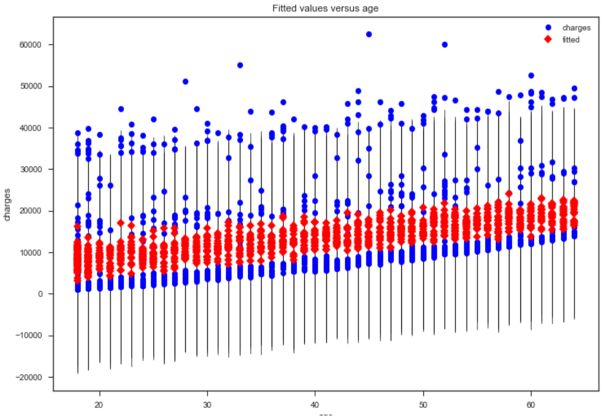


Fig 8: Fitted Values with Age

Further implementation focuses more on actual data analysis[2] rather than judging the dataset. Regression plots can be used to understand the behaviour of the dataset. But before plotting the regression plots, we have to train the dataset. Training the dataset can take from about 20-25 minutes to even few days.

The training of dataset takes time when there are large number of entries and large number of attributes attached to them. The time complexity of such tasks looks bad from the outside, but taking into account the large nature of dataset, and amount of processes done, the algorithm works pretty well.

The regression plot for age, fitted between Y and X gives an interesting perspective as to how the dataset is behaving. On the y axis, we have charges, and on the X axis we have fitted age. On close inspection we can notice that as the age increases, the charges also increase, and the band of fitted value is a uniform one. Also, we notice some outliers that are present in the data. These outliers are clearly visible at random points in the graph. These outliers do not conform to any rule. Regression plots in python are present in the seaborn package. This helps in visualizing the linear relationship between the parameters in discussion.

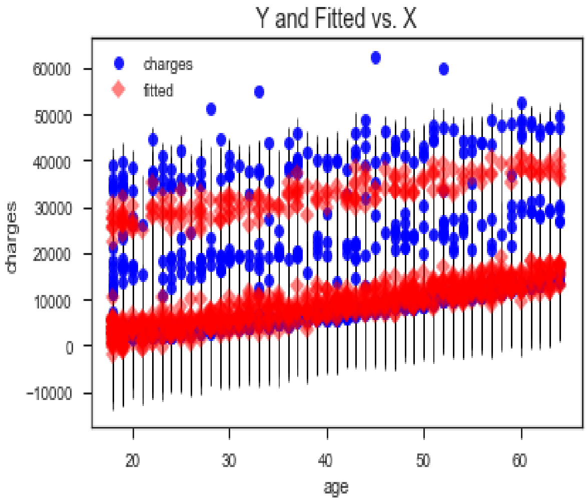


Fig 9: Y and Fitted vs X

The next regression plot is between the residuals and age.

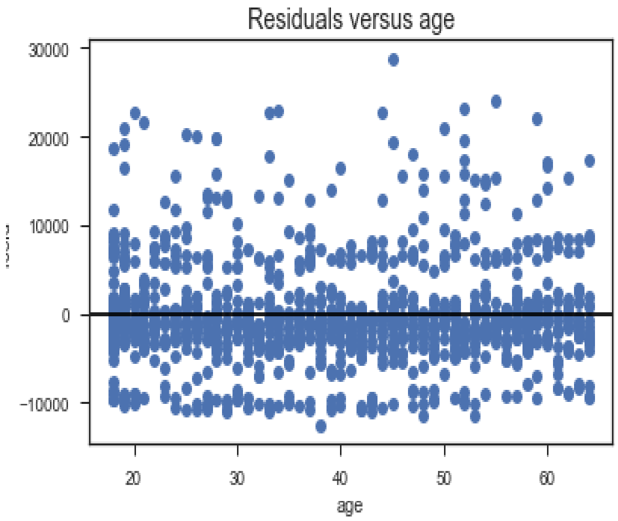


Fig 10: residuals vs Age

We can see the residuals plotted against age in the above figure. It depicts how close a lot of the residuals are.

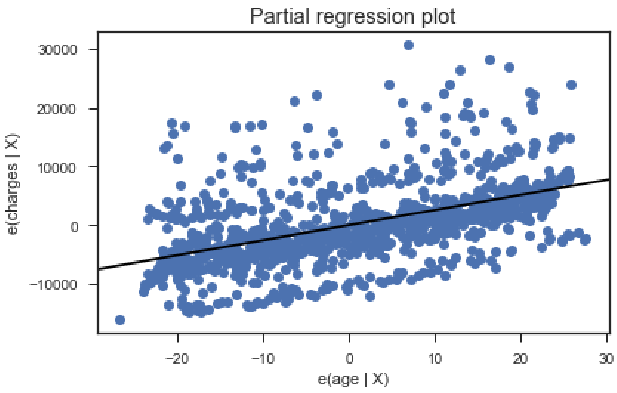


Fig 11: Partial Regression Plot

The partial regression plot shows how the model has varied due to the addition of another regressor, or independent variable. It also helps in depicting the relationship between independent variables.

The final regression plot is the CCPR Plot. CCPR stands for Component-Component plus Residual. The CCPR plot provides a way to judge the effect of one of the independent variables called the regressor, by also considering the effects of other regressor variables. If the variable is highly correlated with other variables, then the output will be inadmissible.

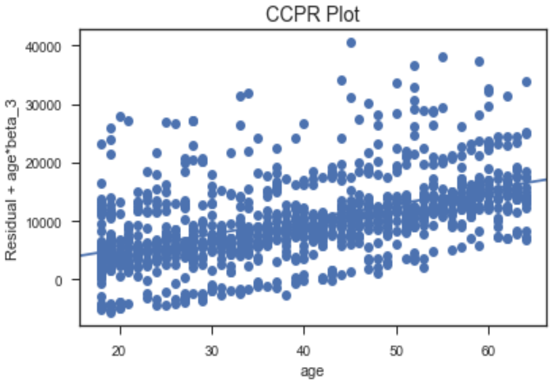


Fig 12: CCPR Plot

VI. CONCLUSION

We have various different types of Insurance. Even though there is admittedly less knowledge about what they are and how to use them, people are gaining knowledge about it every day. Insurance is a safety net for people who happen to own some very valuable commodity, not least their life. Insurance companies make profit when their customers pay them money on a regular basis for some product, and then they never get the opportunity to claim the insurance. The whole insurance industry works under the working principle that not all customers would face a situation where they are forced to claim the insurance. Lot of frauds happen related to insurance, people claim to be dead to claim the life insurance money, and to do so, they fake their own death. Situations also arise where the beneficiaries of the insurance end up harming the policy holder, to get the money quickly. In this paper we are proposing to use, linear regression analysis[4], to predict the patient charges and forecast the insurance values of patients.

In the available dataset used to train the model, we have over 1300 patients and 7 attributes for them each. Each of these attributes is in a way some sort of information about the patient. We will be using python and its varied set of packages purely available for data science, to plot graphs and have an interactive illustration. The results or of expected accuracy good enough to be used directly by any industry.

In the future various enhancements can be made to this model. With the evolution of Artificial Intelligence, we will get more and more accurate unsupervised learning techniques to train our machine to learn from the available dataset. This analysis could also be used for some other project, but with the required changes.

REFERENCES

1. B. S. Todd, R Stamper. On Formal specification of a rule-based expert system.1992
2. Ying Yu, Min Li, Liangliang Liu, Yaohang Li, Jianxin Wang. Clinical big data and deep learning: Applications, challenges, and future outlooks 2019.
3. Emir Slanjankic, Haris Balta, Adil Joldic, Alsa Cvitkovic, Djenan Heric, Emir Veledar. Data mining techniques and SAS as a tool for graphical presentation of principal components analysis and disjoint cluster analysis results. 2019
4. Priya Stephen, Suresh Jaganathan. Linear regression for pattern recognition 2014.
5. Chiat-Pin Tay, Sharmili Roy, Kim-Hui Yap .Multitask Person Re-Identification using Homoscedastic Uncertainty Learning 2019.
6. Mrityunjay Sharma, Suman Saha. Graph based approach for minimum multicollinearity highly accurate regression model explaining maximum variability 2014.
7. Yuke Qiu, Liu Liu, Xin Lai , Yuwen Qiu. An Online Test for Goodness-of-Fit in Logistic Regression Model 2019
8. S.P. Smith. A test to determine the multivariate normality of a data set 1998.
9. Alper Sarikaya , Michael Gleicher .Scatterplots: Tasks, Data, and Designs 2017.

AUTHORS PROFILE

Jayakrishna Natarajan was born in Coimbatore, Tamil Nadu, India in 1999. He is expected to complete his Bachelor of Technology (B-Tech) in the branch of Computer Science & Engineering from SRM Institute of Science & Technology in 2020. His research interests include data science, deep learning with neural networks and machine learning.

Krishanu Das was born in Guwahati, Assam, India in 1998. He is expected to complete his Bachelor of Technology (B-Tech) in the branch of Computer Science & Engineering from SRM Institute of Science & Technology in 2020. His research interests include cyber security and ethical hacking.

Ronak Harish Patil was born in Dhule, Maharashtra, India in 1997. He is expected to complete his Bachelor of Technology (B-Tech) in the branch of Computer Science and Engineering from SRM Institute of Science and Technology in 2020. His research interest includes Machine Learning, Deep Learning and Android App development.

Mr. R B Sarooraj, born in Tamil Nadu, completed his Bachelor of Engineering in the branch of Information Technology from Anna University, Chennai 2009. He completed his Master of Engineering n Computer and Communication Engineering from Anna University, Chennai in 2012. He's working as Assistant Professor in the Department of Computer Science and Engineering at SRM Institute of Science and Technology, Chennai from 2013 till date. His research interests include Data Mining and Analysis of Algorithms.

Retrieval Number: E2098039520/2020©BEIESP
DOI: 10.35940/ijitee.E2098.039520
Journal Website: www.ijitee.org

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication